सत्यमेव जयते

I.A.R I.6.
GIP NLK—H 3 I.A.R.I.—10-5-55—15,000

THE ANNALS
of
MATHEMATICAL STATISTICS

THE ANNALS OF MATHEMATICAL STATISTICS IS AFFILIATED WITH THE AMERICAN STATISTICAL ASSOCIATION AND IS DEVOTED TO THE THEORY AND APPLICATION OF MATHEMATICAL STATISTICS

Volume VII, 1936

PUBLISHED QUARTERLY
ANN ARBOR, MICHIGAN

ON THE FREQUENCY FUNCTION OF xy

By Cecil C. Craig

Given the distribution function of x and y, what can be said of the distribution of the product xy? The author has had two inquiries during the last two years, one from an investigator in business statistics and the other from a psychologist, concerning the probable error of the product of two quantities, each of known probable error. There seems to be very little in the literature of mathematical statistics on this question.

If x and y are independent and are each distributed according to the same normal frequency law, it is well known that the distribution function of

$$\bar{z} = \frac{x - m_x}{\sigma_x} \cdot \frac{y - m_y}{\sigma_y}$$

is

$$\frac{1}{\pi} K_0(\bar{z}) \text{ ,}^1$$

in which $K_0(\bar{z})$ is the Bessel function of the second kind of a purely imaginary argument of zero order.[2] If x and y are independent and are each distributed according to a logarithmic normal frequency law, it has been pointed out that the product, $(x - a)\ (y - b)$, in which a and b are the upper (or lower) limits of the range for x and y respectively, is distributed according to a law of the same type.[3] In both cases the special choice of origins greatly simplifies the problem.

In the present discussion it will be assumed that x and y are distributed normally. It will appear that the distribution of xy is a function of r_{xy}, the coefficient of correlation between x and y, and of the parameters,

$$\rho_1 = \frac{m_1}{\sigma_1} = \frac{m_x}{\sigma_x} \quad \text{and} \quad \rho_2 = \frac{m_2}{\sigma_2} = \frac{m_y}{\sigma_y} \text{ ,}$$

which are proportional to the reciprocals of the coefficients of variation. The chief difficulty arises when ρ_1 and ρ_2 are small so that zero values of xy occur

[1] J. Wishart and M. S. Bartlett: The Distribution of Second Order Moment Statistics in a Normal System; Proceedings of the Cambridge Philosophical Society, Vol. XXVIII (1932), pp. 455–459.

[2] G. N. Watson: A Treatise on the Theory of Bessel Functions; Cambridge University Press (1922), p. 78.

[3] P. T. Yuan: On the Logarithmic Frequency Distribution and the Semi-logarithmic Frequency Surface; Annals of Mathematical Statistics, Vol. 4 (1933), pp. 46, 47.

for values of x and y well within their respective ranges of variation. (If ρ_1 and ρ_2 are large, practically one may exclude zero values of x and y from consideration. The author hopes to present an investigation of this case soon.) It is the object of the present paper to study the rather unusual frequency function that arises in this situation. It will first be assumed that x and y are independent ($r_{xy} = r = 0$). Then it will be shown that the distribution function when $r \neq 0$ is readily derived from that arrived at in the special case.

We can find the moment generating function of xy without difficulty. We have,

$$M_{xy}(\vartheta) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}-\frac{(y-m_2)^2}{2\sigma_2^2}} e^{xy\vartheta}\,dx\,dy$$

$$= \frac{e^{[(\sigma_1^2 m_2^2+\sigma_2^2 m_1^2)\vartheta^2+2m_1m_2\vartheta]/2(1-\sigma_1^2\sigma_2^2\vartheta^2)}}{(1-\sigma_1^2\sigma_2^2\vartheta^2)^{1/2}}.$$

Setting, for convenience,

$$z = \frac{xy}{\sigma_1\sigma_2},$$

this can be written,

$$M_z(\vartheta) = \frac{e^{[(\rho_1^2+\rho_2^2)\vartheta^2+2\rho_1\rho_2\vartheta]/2(1-\vartheta^2)}}{(1-\vartheta^2)^{1/2}}. \tag{1}$$

This choice of variable and of parameters will be adhered to in the sequel.

On expanding $\log M_z(\vartheta)$ in powers of ϑ, we get for the semi-invariants (of Thiele),

$$\begin{aligned}\lambda_{2k+1:z} &= (2k+1)!\,\rho_1\rho_2, \qquad k = 0, 1, 2, \cdots\\ \lambda_{2k:z} &= \frac{(2k)!}{2}(\rho_1^2+\rho_2^2) + (2k-1)!, \qquad k = 1, 2, \cdots.\end{aligned} \tag{2}$$

These give for the mean and variance of xy,

$$M_{xy} = m_1m_2$$

$$\sigma_{xy}^2 = \sigma_1^2m_2^2 + \sigma_2^2m_1^2 + \sigma_1^2\sigma_2^2.$$

For the standard semi-invariants of z (or of xy), we have,

$$\xi_{2k+1:z} = \frac{\lambda_{2k+1:z}}{\lambda_{2:z}^{\frac{2k+1}{2}}} = \frac{(2k+1)!\,\rho_1\rho_2}{(\rho_1^2+\rho_2^2+1)^{\frac{2k+1}{2}}},$$

$$\xi_{2k:z} = \frac{\lambda_{2k:z}}{\lambda_{2:z}^{k}} = \frac{(2k-1)!\,[k(\rho_1^2+\rho_2^2)+1]}{(\rho_1^2+\rho_2^2+1)^k}$$

Taking,

$$\xi_3 = \frac{6\,\rho_1\rho_2}{(\rho_1^2 + \rho_2^2 + 1)^{3/2}},$$

as a measure of skewness, it is easy to verify that

$$|\,\xi_3\,| \leqq \frac{2}{3}\sqrt{3}\,.$$

For either $\rho_1 = 0$ or $\rho_2 = 0$, the distribution is symmetrical about its mean which then falls at the origin.

For the excess or kurtosis, we have,

$$\xi_4 = \frac{6\,[2(\rho_1^2 + \rho_2^2) + 1]}{(\rho_1^2 + \rho_2^2 + 1)^2} \leqq 6\,.$$

Thus the skewness is never great and becomes small with increasing ρ_1 or ρ_2. The excess also becomes small with increasing ρ_1 or ρ_2, but it can be very large for small values of these parameters, attaining its maximum of 6 for $\rho_1 = \rho_2 = 0$. But, as it will appear below, the distribution function always becomes infinite in a logarithmic manner at the origin. (We have already seen, as must obviously be the case, that moments of all orders exist.) It is to be noted, too, that for any given ρ_1 and ρ_2, ξ_{2k} increases without limit with increasing k, and that the same is true of ξ_{2k+1} if neither ρ_1 nor ρ_2 is zero.

Turning now to the derivation of the actual frequency function of z, we set $w = xy$; then for any given x, $y = w/x$, $dy = \dfrac{dw}{x}$ if $x > 0$, and $dy = -\dfrac{dw}{x}$ if $x < 0$. These values are substituted into $\varphi_1(x)\,\varphi_2(y)dx\,dy$, in which $\varphi_1(x)$ and $\varphi_2(y)$ are the frequency functions of x and y respectively, and the resulting expression is integrated over all values of x, giving for the frequency function of w:

$$F(w) = \frac{e^{-\left(\frac{m_1^2}{2\sigma_1^2} + \frac{m_2^2}{2\sigma_2^2}\right)}}{2\pi\sigma_1\sigma_2}\left[\int_0^\infty \Phi(w, x)\,\frac{dx}{x} - \int_{-\infty}^0 \Phi(w, x)\,\frac{dx}{x}\right]$$

in which,

$$\Phi(w, x) = e^{-(\sigma_2^2x^4 - 2m_1\sigma_2^2x^2 - 2m_2\sigma_1^2wx + \sigma_1^2w^2)/2\sigma_1^2\sigma_2^2x^2}.$$

Again setting $z = \dfrac{xy}{\sigma_1\sigma_2}$, and introducing the parameters ρ_1 and ρ_2, this reduces to,

$$F(z) = \frac{e^{-\frac{(\rho_1^2+\rho_2^2)}{2}}}{2\pi}[\psi_1(z) - \psi_2(z)]\,, \tag{4}$$

in which,

$$\psi_1(z) = \int_0^\infty e^{-\left(\frac{x^2}{2} - \rho_1 x - \rho_2 \frac{z}{x} + \frac{z^2}{2x^2}\right)} \frac{dx}{x}, \tag{5}$$

and $\psi_2(z)$ is the integral of the same function over the interval $(-\infty, 0)$.

Now writing,

$$\psi_1(z) = \int_0^\infty e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} \frac{e^{\rho_1 x + \rho_2 \frac{z}{x}}}{x} dx, \tag{6}$$

we note that

$$\frac{e^{\rho_1 x + \rho_2 \frac{z}{x}}}{x}$$

can be expanded in a Laurent series in powers of x for all values of x except zero. In this expansion the coefficient of x^{r-1}, $r \geqq 1$, is $\frac{\rho_1^r}{r!} \sum_r(\rho_1\rho_2 z)$, in which

$$\sum_r(\rho_1\rho_2 z) = 1 + \frac{\rho_1\rho_2 z}{r+1} + \frac{(\rho_1\rho_2 z)^2}{(r+2)^{(2)}\, 2!} + \frac{(\rho_1\rho_2 z)^3}{(r+3)^{(3)}\, 3!} + \cdots, \tag{7}$$

$$((r+k)^{(k)} = (r+k)(r+k-1)\cdots(r+1)).$$

We may note parenthetically that

$$\frac{\rho_1^r}{r!} \sum_r(\rho_1\rho_2 z) = \left(\frac{\rho_1}{\rho_2 z}\right)^{\frac{r}{2}} I_r(2\sqrt{\rho_1\rho_2 z}),$$

in which $I_r(x)$ is the Bessel function of the first kind with a purely imaginary argument.[4]

The coefficient of x^{-r-1}, $r \geqq 0$, is $\frac{z^r \rho_2^r}{r!} \sum_r(\rho_1\rho_2 z)$.

Setting now,

$$\sum_{n=-\infty}^{\infty} f_n(x) = \frac{e^{\rho_1 x + \rho_2 \frac{z}{x}}}{x},$$

we substitute this series in (6) and seek to justify the expansion it gives for $\psi_1(z)$ obtained by term by term integration. We write,

$$\psi_1(z) = \int_0^1 e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} \sum f_n(x)\, dx + \int_1^\infty e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} \sum f_n(x)\, dx.$$

[4] Watson, loc. cit., p. 77.

For $z > 0$, $\rho_1\rho_2 > 0$, the terms of $\sum f_n(x)$ are all > 0. Then the convergence of

$$\sum \int_0^1 e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} f_n(x)\, dx$$

is sufficient to allow term by term integration in the first integral. In the second integral we observe that $\sum f_n(x)$ converges uniformly in every fixed interval $1 \leqq x \leqq a$. Then term by term integration is permissible here if

$$\sum \int_0^\infty e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} f_n(x)\, dx$$

is convergent.[5] It is evident, then, that it will be sufficient to establish the convergence of

$$\sum \int_0^\infty e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} f_n(x)\, dx\,.$$

If either or both $z < 0$ or $\rho_1\rho_2 < 0$, it will be easily seen that the series involved are still absolutely convergent which is sufficient.

Now using the definition of the Bessel function of a purely imaginary argument of the second kind,

$$K_\nu(z) = \frac{1}{2}\left(\frac{z}{2}\right)^\nu \int_0^\infty e^{-\tau - \frac{z^2}{4\tau}} \frac{d\tau}{\tau^{\nu+1}}\,,^6$$

it is easy to derive the relation,

$$K_{\frac{n-1}{2}}(z) = z^{\frac{n-1}{2}} \int_0^\infty e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} \frac{dx}{x^n}\,.$$

Remembering that $K_\nu(z) = K_{-\nu}(z)$, we have for our expansion,

$$\psi_1(z) = \sum\nolimits_0 K_0 + (\rho_1 + \rho_2)\, z^{\frac{1}{2}} \sum\nolimits_1 K_{\frac{1}{2}} + (\rho_1^2 + \rho_2^2)\, \frac{z}{2!} \sum\nolimits_2 K_1$$

$$+ (\rho_1^3 + \rho_2^3)\, \frac{z^{\frac{3}{2}}}{3!} \sum\nolimits_3 K_{\frac{1}{2}} + \cdots$$

in which the argument for all the $\sum$-functions is $\rho_1\rho_2\, z$, and for all the K-functions is z.

---

[5] T. J. I'a Bromwich: An Introduction to the Theory of Infinite Series; Macmillan & Co., London, 2nd edition (1926), pp. 496 and 500.

[6] Watson, loc. cit., pp. 78 and 183.

But we may as well add to this the expansion of $-\psi_2(z)$, which may be written,

$$\int_0^\infty e^{-\frac{x^2}{2}-\frac{z^2}{2x^2}} \frac{e^{-\rho_1 x-\rho_2\frac{z}{x}}}{x}\,dx\,,$$

and obtain the expansion,

$$F(z)=\frac{e^{-\frac{\rho_1^2+\rho_2^2}{2}}}{\pi}\left[\sum_0 K_0+(\rho_1^2+\rho_2^2)\frac{z}{2!}\sum_2 K_1+(\rho_1^4+\rho_2^4)\frac{z^2}{4!}\sum_4 K_2+\cdots\right],$$

the convergence of which we will examine. But it must be noted that the terms arising from the expansion of

$$\frac{e^{\rho_1 x+\rho_2\frac{z}{x}}}{x} \quad \text{and} \quad \frac{e^{-\rho_1 x-\rho_2\frac{z}{x}}}{x}$$

which contribute to the expansion of $F(z)$ as just written are those of the forms,

$$\frac{\rho_1^{2i}}{(2i)!}\sum_{2i} \quad \text{and} \quad \frac{\rho_2^{2i}\,z^{2i}}{(2i)!}\sum_{2i}\,.$$

Hence the expansion as written is valid in any case only for $z>0$. For $z\gtrless 0$, we may write however,

$$F(z)=\frac{e^{-\frac{\rho_1^2+\rho_2^2}{2}}}{\pi}\left[\sum_0 K_0+(\rho_1^2+\rho_2^2)\frac{|z|}{2!}\sum_2 K_1+(\rho_1^4+\rho_2^4)\frac{z^2}{4!}\sum_4 K_2 + (\rho_1^6+\rho_2^6)\frac{|z|^3}{6!}\sum_6 K_3+\cdots\right], \tag{8}$$

in which the arguments for the $\sum$ and K-functions are the same as before.

Let us consider now the question of the convergence of (8), first in the case that $z>0$. We set

$$c_\nu=\frac{z^\nu K_\nu}{\nu!}\Big/\frac{z^{\nu-1}K_{\nu-1}}{(\nu-1)!}\,.$$

Then from the relation,

$$K_{\nu-1}-K_{\nu+1}=-\frac{2\nu}{z}K_\nu\,,^{7} \tag{9}$$

we readily derive,

$$\frac{z^2}{(\nu+1)^{(2)}}=c_\nu\left(c_{\nu+1}-\frac{2\nu}{\nu+1}\right).$$

[7] Watson, loc. cit., p. 79.

For $z > 0$, the left hand member and c_ν are both > 0. Thus

$$c_{\nu+1} - \frac{2\nu}{\nu + 1} > 0 .$$

Then let

$$c_{\nu+1} = \frac{2\nu}{\nu + 1} + \delta_{\nu+1} , \qquad \delta_{\nu+1} > 0 ,$$

and we have,

$$\frac{z^2}{(\nu + 1)^{(2)}} > \left(2 - \frac{2}{\nu}\right) \delta_{\nu+1} = 2\,\delta_{\nu+1} - \frac{2\,\delta_{\nu+1}}{\nu} .$$

It is evident from this that for a given $z > 0$, a ν_0 exists such that $c_\nu < 3$ for $\nu \geqq \nu_0$.

Further since

$$\sum_r \leqq e^{|\rho_1 \rho_2 z|}$$

the convergence sought follows for $z > 0$. Since K is an even function of z, it is easy to see that (8) is also convergent for $z < 0$. For $z = 0$, the first term possesses a logarithmic discontinuity at the origin.

To calculate ordinates of $F(z)$ there are fairly extensive tables available in Watson's treatise already referred to. These tables may be readily extended by means of the asymptotic formula for $K(z)$ for larger values of z, and by means of (9) for larger values of ν. One can rapidly build up tables of $\sum_r(x)$ by means of the easily obtained recursion formula,

$$\sum_r(x) = \sum_{r+1}(x) + \frac{x}{(r + 2)^{(2)}} \sum_{r+2}(x) .$$

It is unfortunately true that the expansion found for $F(z)$ is very slowly convergent for large values of ρ_1 and ρ_2.

At the end of this paper are shown three charts of $F(z)$ with the tables of ordinates from which they were made by way of illustrating what such curves look like. (On the second for comparison the broken line is the normal curve of error.)

For $\rho_1 = \rho_2 = r = 0$, we have simply the known result,

$$F(z) = \frac{1}{\pi} K_0(z) .$$

For $\rho_1 = 1$, $\rho_2 = r = 0$, the curve is symmetrical about its mean (and the origin). Here every $\sum$-function is unity.

For the case in which $\rho_1 = \rho_2 = \frac{1}{2}$, $r = 0$, I first constructed tables of $\sum_i(x)$ for $x = \pm 0.025, \pm 0.05, \pm 0.1$, and by intervals of 0.1 to ± 3.0 for $i = 0, 1, \cdots, 20$. Values of $\sum_0(x)$ and $\sum_2(x)$ for $x = 3.2$ and 3.4 were also used. Not more than five terms of (8) were required to obtain values of $F(z)$ accurate

to five places of decimals. This distribution curve is skew with $M_s = 0.25$ and $\xi_{3:s} = \frac{\sqrt{6}}{3}$.

The curves are plotted in standard units with unit total area ($\sigma_s = \sqrt{\rho_1^2 + \rho_2^2 + 1}$). The tables of ordinates are given both in units of $z = \frac{xy}{\sigma_1\sigma_2}$ and of $t = \frac{z - m_s}{\sigma_s}$.

Turning now to the case in which $r \neq 0$, after some computation, we have for the moment generating function,

$$M_s(\vartheta) = \frac{e^{\frac{(\rho_1^2+\rho_2^2-2r\rho_1\rho_2)\vartheta^2+2\rho_1\rho_2\vartheta}{2[1-(1+r)\vartheta][1+(1-r)\vartheta]}}}{\sqrt{[1-(1+r)\vartheta][1+(1-r)\vartheta]}}. \tag{10}$$

As a check on this result, if we set $r = 1$ and $\rho_1 = \rho_2 = \rho$ in it we get,

$$M_{\frac{x^2}{\sigma^2}}(\vartheta) = \frac{e^{\frac{\rho^2\vartheta}{1-2\vartheta}}}{\sqrt{1-2\vartheta}},$$

which may be readily verified to be the moment generating function of $\frac{x^2}{\sigma^2}$ if x is distributed normally with mean m and variance σ^2 $\left(\rho = \frac{m}{\sigma}\right)$.

To obtain the semi-invariants of z in this case, on expanding $\log M_s(\vartheta)$ in powers of ϑ, setting

$$a = \rho_1^2 + \rho_2^2 - 2\rho_1\rho_2 r, \quad b = 2\rho_1\rho_2, \quad c = 1 + r, \quad \text{and} \quad d = 1 - r,$$

we have,

$$\begin{aligned}\log M_s(\vartheta) &= \frac{a\vartheta^2 + b\vartheta}{2}(1 - c\vartheta)^{-1}(1 + d\vartheta)^{-1} \\ &\quad - \frac{1}{2}[\log(1 - c\vartheta) + \log(1 + d\vartheta)] \\ &= \frac{a\vartheta^2 + b\vartheta}{4}[2 + (c^2 - d^2)\vartheta + (c^3 + d^3)\vartheta^2 + (c^4 - d^4)\vartheta^3 + \cdots] \\ &\quad + \frac{1}{2}\left[(c - d)\vartheta + (c^2 + d^2)\frac{\vartheta^2}{2} + (c^3 - d^3)\frac{\vartheta^3}{3} + \cdots\right],\end{aligned} \tag{11}$$

from which we derive,

$$\begin{aligned}\lambda_{n:s} &= \frac{n!}{4}[\{c^{n-1} - (-d)^{n-1}\}a + \{c^n - (-d)^n\}b] \\ &\quad + \frac{(n-1)!}{2}\{c^n + (-d)^n\}.\end{aligned} \tag{12}$$

In particular,

$$\lambda_{1:s} = \frac{b}{2} + \frac{c-d}{2} = \rho_1\rho_2 + r$$

$$\lambda_{2:s} = a + \frac{c^2-d^2}{2}\cdot b + \frac{c^2+d^2}{2} = \rho_1^2 + \rho_2^2 + 2\rho_1\rho_2 r + (1+r^2)$$

$$\lambda_{3:s} = \frac{3}{2}\left[(c^2-d^2)\,a + (c^3+d^3)\,b\right] + c^3 - d^3 \tag{13}$$

$$= 6\left[(\rho_1^2+\rho_2^2)\,r + \rho_1\rho_2\,(1+r^2)\right] + 2\,r\,(3+r^2)$$

$$\lambda_{4:s} = 6\left[(c^3+d^3)\,a + (c^4-d^4)\,b\right] + 3\,(c^4+d^4)$$

$$= 12\,(\rho_1^2+\rho_2^2)\,(1+3\,r^2) + 24\,\rho_1\rho_2 r\,(3+r^2) + 6\,(1+6r+r^4)\,.$$

Noting that

$$\frac{\partial a}{\partial r} = -\,b\,, \qquad \frac{\partial b}{\partial r} = 0\,, \qquad \frac{\partial c}{\partial r} = 1\,, \qquad \frac{\partial d}{\partial r} = -\,1\,,$$

one can easily demonstrate what seems to be a rather striking property of these semi-invariants, viz.,

$$\frac{\partial \lambda_{n:s}}{\partial r} = n(n-1)\,\lambda_{n-1:s}\,. \tag{14}$$

To gain a notion of the magnitude of the skewness and excess in this case, we form,

$$\xi_{3:s} = \frac{\lambda_{3\,s}}{\lambda_{2:s}^{\frac{3}{2}}} \quad \text{and} \quad \xi_{4:s} = \frac{\lambda_{4:s}}{\lambda_{2:s}^2}\,.$$

In view of the above property,

$$\frac{\partial \xi_3}{\partial r} = \frac{6\,\lambda_2^2 - 3\,\lambda_3\lambda_1}{\lambda_2^{\frac{5}{2}}}\,.$$

The denominator of this fraction is always > 0. The numerator, after some reduction, can be written,

$$6\left[\rho_1^4 + \rho_2^4 - \rho_1^2\rho_2^2\,(1-r^2) + (\rho_1^2+\rho_2^2)\,(2-r^2)\right. \tag{15}$$
$$\left. + (\rho_1^2+\rho_2^2-2)\,\rho_1\rho_2 r + 1 - r^2\right]\,.$$

The first two terms taken together, the third, and the last are all obviously > 0. The term,

$$(\rho_1^2+\rho_2^2-2)\,\rho_1\rho_2 r$$

has its maximum value for $|r| = 1$. But for $r = 1$, (15) becomes,

$$\rho_1^4 + \rho_2^4 + \rho_1\rho_2(\rho_1^2 + \rho_2^2) + (\rho_1 - \rho_2)^2,$$

and for $r = -1$, it is,

$$\rho_1^4 + \rho_2^4 - \rho_1\rho_2(\rho_1^2 + \rho_2^2) + (\rho_1 + \rho_2)^2,$$

both of which expressions are easily seen to be > 0.

Thus (15) is always positive and the maximum value of $\xi_{3:s}$ is attained for $r = 1$, the minimum value for $r = -1$. These values are respectively,

$$\frac{6(\rho_1 + \rho_2)^2 + 8}{[(\rho_1 + \rho_2)^2 + 2]^{\frac{3}{2}}} \quad \text{and} \quad \frac{-6(\rho_1 - \rho_2)^2 - 8}{[(\rho_1 - \rho_2)^2 + 2]^{\frac{3}{2}}},$$

the absolute value of either being $\leqq 2\sqrt{2}$, which is attained in the first case for $\rho_1 = -\rho_2$ and in the second for $\rho_1 = \rho_2$. It is seen that for high correlation between x and y the skewness of xy can be quite large.

For the excess, we see that

$$\xi_{4.s} = \frac{\lambda_{4\cdot s}}{\lambda_{2.s}^2}$$

attains a value of 12 when $\rho_1 = -\rho_2$, $r = 1$ or when $\rho_1 = \rho_2$, $r = -1$. Since this is such an extraordinary value it does not seem worth while to carry out the extended computation that seems to be required to verify one's surmise that this is the maximum of the absolute value.

Now, to derive the frequency function we proceed as before. We set $z = \frac{xy}{\sigma_1\sigma_2}$ and then

$$F(z) = I_1(z) - I_2(z),$$

in which,

$$I_1(z) = \frac{1}{2\pi\sqrt{1 - r^2}} \int_0^\infty e^{-\frac{1}{2(1-r^2)}\left[(x-\rho_1)^2 - 2r(x-\rho_1)\left(\frac{z}{x}-\rho_2\right) + \left(\frac{z}{x}-\rho_2\right)^2\right]} \frac{dx}{x},$$

and $I_2(z)$ is the integral of the same function over the interval $(-\infty, 0)$.

We can write $I_1(z)$:

$$\frac{e^{-\frac{\rho_1^2 - 2r\rho_1\rho_2 + \rho_2^2}{2(1-r^2)} + \frac{rz}{1-r^2}}}{2\pi\sqrt{1 - r^2}} \int_0^\infty e^{-\frac{1}{2(1-r^2)}\left(x^2 + \frac{z^2}{x^2}\right) + \frac{1}{1-r^2}\left[(\rho_1 - r\rho_2)x + (\rho_2 - r\rho_1)\frac{z}{x}\right]} \frac{dx}{x}.$$

Setting,

$$\frac{x}{\sqrt{1 - r^2}} = u \quad \text{and} \quad \frac{z}{1 - r^2} = \zeta,$$

this becomes,

$$(16)\qquad \frac{\sqrt{1-r^2}\,e^{-\frac{\rho_1^2-2r\rho_1\rho_2+\rho_2^2}{2(1-r^2)}+\frac{r\zeta}{(1-r^2)^2}}}{2\pi}\times\int_0^\infty e^{-\frac{1}{2}\left(u^2+\frac{\zeta^2}{u^2}\right)}\,e^{\frac{\rho_1-r\rho_2}{\sqrt{1-r^2}}u+\frac{\rho_2-r\rho_1}{\sqrt{1-r^2}}\frac{\zeta}{u}}\frac{du}{u}.$$

But on writing,

$$\frac{\rho_1-r\rho_2}{\sqrt{1-r^2}}=R_1\quad\text{and}\quad\frac{\rho_2-r\rho_1}{\sqrt{1-r^2}}=R_2,$$

the integral in the last expression is of the same form as the $\psi_1(z)$ in the uncorrelated case. It is evident, then, that the distribution function of ζ can be written,

$$\frac{\sqrt{1-r^2}}{\pi}\,e^{-\frac{\rho_1^2-2r\rho_1\rho_2+\rho_2^2}{2(1-r^2)}}\,e^{\frac{r\zeta}{(1-r^2)^2}}\Big[\sum_0(R_1R_2\zeta)K_0(\zeta)$$

$$(17)\qquad+(R_1^2+R_2^2)\frac{|\zeta|}{2!}\sum_2(R_1R_2\zeta)K_1(\zeta)+(R_1^4+R_2^4)\frac{\zeta^2}{4!}\sum_4(R_1R_2\zeta)K_2(\zeta)$$

$$+(R_1^6+R_2^6)\frac{|\zeta|^3}{6!}\sum_6(R_1R_2\zeta)K_3(\zeta)+\cdots\Big],$$

and is essentially of the form of $F(z)$, reached when $r=0$, multiplied by an exponential function.

Frequency curves for xy (in standard units) are given in Fig. 1, Fig. 2 and Fig. 3.

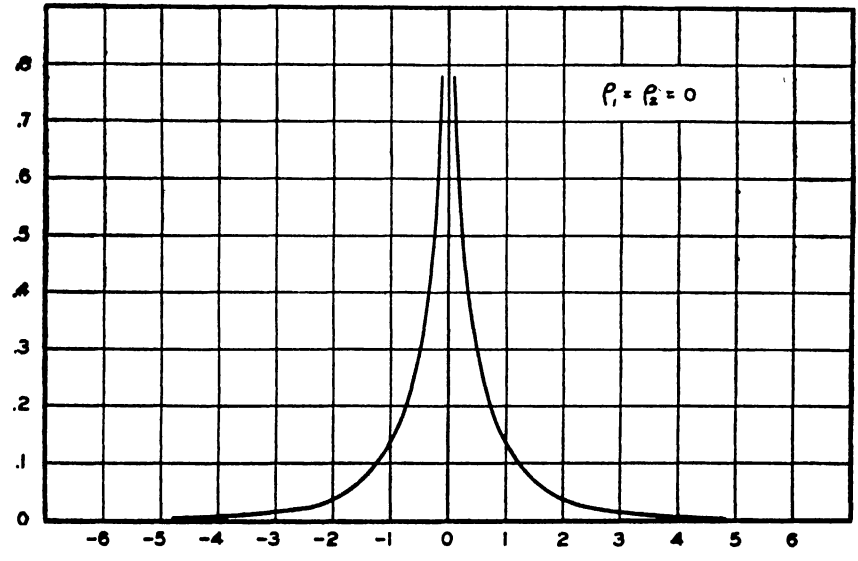


FIG. 1

TABLES OF ORDINATES OF THE DISTRIBUTION FUNCTIONS, $F(z)$ AND $F(t)$

For $\rho_1 = \rho_2 = 0, r = 0$ (Curve is symmetrical with respect to origin) $M_s = 0, \sigma_s = 1$		$\rho_1 = 1, \rho_2 = 0, r = 0$ (Curve is symmetrical with respect to origin) $M_s = 0, \sigma_s = \sqrt{2}$			
$z = t$	$F(z) = F(t)$	z	$F(z)$	t	$F(t)$
0.1	0.77256	0.1	0.58215	0.07	0.82328
0.2	.55790	0.2	.44891	.14	.63485
0.3	.43887	0.3	.37159	.21	.52551
0.4	.35477	0.4	.31736	.28	.44882
0.5	.29425	0.5	.27593	.35	.39023
0.6	0.24749	0.6	0.24270	0.42	0.34323
0.7	.21025	0.7	.21519	.49	.30432
0.8	.17996	0.8	.19193	.57	.27143
0.9	.15493	0.9	.17195	.64	.24318
1.0	.13402	1.0	.15460	.71	.21863
1.2	0.10138	1.2	0.12595	0.85	0.17812
1.4	.07756	1.4	.10340	0.99	.14623
1.6	.05983	1.6	.08533	1.13	.12068
1.8	.04645	1.8	.07069	1.27	.09997
2.0	.03625	2.0	.05873	1.41	.08306
2.4	0.02235	2.4	0.04078	1.70	0.05767
2.8	.01395	2.8	.02846	1.98	.04025
3.2	.00878	3.2	.01992	2.26	.02818
3.6	.00557	3.6	.01397	2.55	.01976
4.0	.00355	4.0	.00981	2.83	.01388
4.8	0.00146	4.8	0.00485	3.39	0.00685
5.6	.00061	5.6	.00239	3.96	.00338
6.4	.00026	6.4	.00118	4.53	.00167
7.2	.00011	7.2	.00058	5.09	.00082
8.0	.00005	8.0	.00029	5.66	.00040
9.0	0.00002	9.0	0.00012	6.36	0.00017
10.0	.00001	10.0	.00005	7 07	.00007
		11.0	.00002	7.78	.00003
		12.0	.00001	8.49	.00001

$$\rho_1 = \rho_2 = \tfrac{1}{2}, r = 0$$

$$M_s = 0.25, \sigma_s = \frac{\sqrt{6}}{2}.$$

z	$F(z)$	t	$F(t)$
-9.6	0.00001	−8.04	0.00001
-8.8	.00002	−7.39	.00002
-8.0	0.00004	−6.74	0.00005
-7.2	.00010	−6.08	.00012
-6.4	.00023	−5.43	.00028
-5.6	.00054	−4.78	.00066
-4.8	.00128	−4.12	.00157
-4.0	0.00311	−3.47	0.00381
-3.6	.00488	−3.14	.00598
-3.2	.00769	−2.82	.00942
-2.8	.01221	−2.49	.01495
-2.4	.01954	−2.16	.02393
-2.0	0.03165	−1.84	0.03876
-1.6	.05213	−1.51	.06384
-1.2	.08809	−1.18	.10788
-0.8	.15568	−0 86	.19066
-0.4	.30423	−0.53	.37259
-0.2	0.47388	−0.37	0.58036
-0.1	.64994	−0.28	.79598
0.1	0.68106	−0.12	0.83409
0.2	.51947	−0.04	.63619
0.4	0.36322	0.12	0.44484
0.8	.21768	.45	.26659
1.2	.14230	.78	.17427
1.6	.09621	1.10	.11783
2.0	.06614	1.43	.08100
2.4	0.04589	1.76	0.05620
2.8	.03201	2.08	.03920
3.2	.02241	2.41	.02745
3.6	.01571	2.74	.01924
4.0	.01103	3.06	.01351

$$\rho_1 = \rho_2 = \tfrac{1}{2}, r = 0$$

$$M_z = 0.25, \sigma_z = \frac{\sqrt{6}}{2}.$$

z	$F(z)$	t	$F(t)$
4.8	0.00545	3.72	0.00667
5.6	.00269	4.36	.00329
6.4	.00133	5.02	.00163
7.2	.00065	5.67	.00080
8.0	.00032	6.33	.00039
8.8	0.00016	6.98	0.00020
9.6	.00008	7.63	.00010
10.4	.00004	8.29	.00005
11.2	.00002	8.94	.00002
12.0	.00001	9.59	.00001

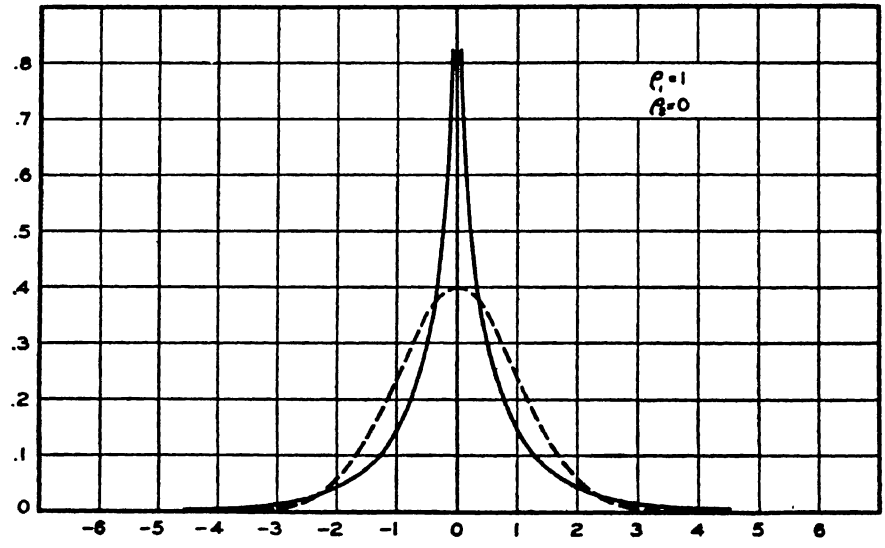


FIG. 2

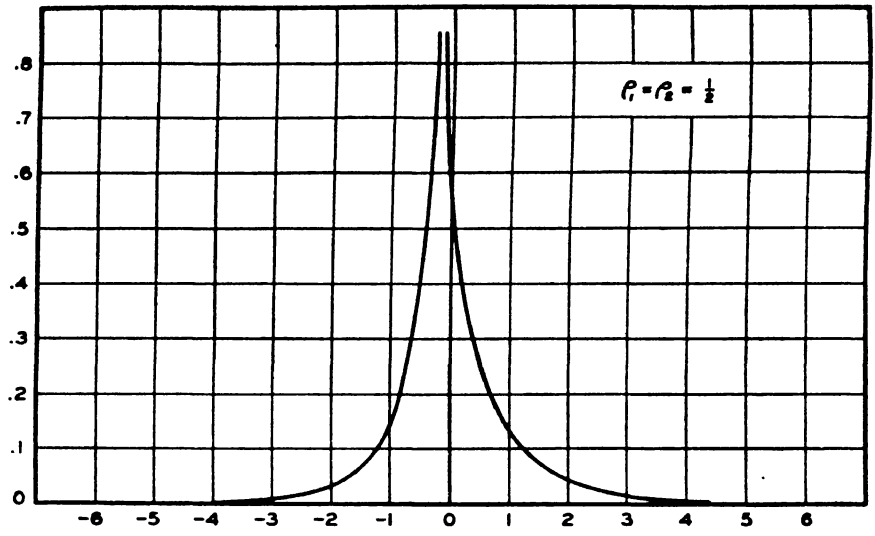


FIG. 3

UNIVERSITY OF MICHIGAN.

A NEW EXPOSITION AND CHART FOR THE PEARSON SYSTEM OF FREQUENCY CURVES

By Cecil C. Craig

In the course of some years of teaching classes in mathematical statistics, the author has expanded the treatment of the Pearson system of frequency functions begun in the Handbook of Mathematical Statistics[1] into an exposition that he believes possesses marked advantages in unity, clarity, and elegance. This is accomplished by expressing the variable in standard units throughout and by making the two parameters $\alpha_3(\alpha_3^2 = \beta_1,\ \alpha_4 = \beta_2$ in Pearson's notation) and

$$\delta = \frac{2\alpha_4 - 3\alpha_3^2 - 6}{\alpha_4 + 3}$$

fundamental in the discussion. The various formulae that arise are obtained directly and in a uniform manner and are relatively simple in form and easy to use. The criteria for the different members of the system of functions are expressed very simply in terms of α_3 and δ and the chart corresponding to the extension of the Rhind diagram given by Pearson[3] takes on a strikingly simple form.

Following the beginning made in the Handbook, the system of Pearson frequency functions are to be found among the solutions of the differential equation

$$\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{b_0 + b_1 t + b_2 t^2}. \tag{1}$$

For those solutions $y = f(t)$ for which,

$$(b_0 + b_1 t + b_2 t^2)\, t^n f(t) \Big]_{t=r}^{s} = 0,$$

[1] H. L. Rietz, Editor-in-Chief; Houghton-Mifflin Co., Boston (1924). See the chapter on Frequency Curves by H. C. Carver.

[2] The notation used is that of the Handbook, loc. cit., to which reference will be frequently made. The discussion of Robert Henderson, "Frequency Curves and Moments," Transactions of the Actuarial Society of America, Vol. VIII (1904), pp. 30–41, also proceeds along very similar lines, although Professor Carver was quite unaware of it when he wrote his chapter in the Handbook. The notation of the Handbook seems preferable however.

[3] Karl Pearson: Mathematical Contributions to the Theory of Evolution, XIX. Second Supplement to a Memoir on Skew Variation; Proc. Roy. Soc., A. Vol. 216 (1916), plate opposite p. 456.

if r and s are the extremes of the range of variation for t, and for which the first $n + 1$ moments over this range exist, the recursion formula for moments,

$$\alpha_n a + n\alpha_{n-1}b_0 + (n + 1)\alpha_n b_1 + (n + 2)\alpha_{n+1}b_2 = \alpha_{n+1}, \tag{2}$$

can be derived. Then setting $n = 0, 1, 2, 3$ we get the following expressions for the parameters, a, b_0, b_1, b_2 in terms of α_3 and δ:

$$\begin{aligned} a &= -\frac{\alpha_3}{2(1 + 2\delta)}, & b_1 &= \frac{\alpha_3}{2(1 + 2\delta)} \\ b_0 &= \frac{2 + \delta}{2(1 + 2\delta)} & b_2 &= \frac{\delta}{2(1 + 2\delta)} \end{aligned} \tag{3}$$

valid except when $\delta = -\frac{1}{2}$. Below note will be taken of those solutions for which the conditions imposed in deriving (2) are not satisfied. The case in which $\delta = -\frac{1}{2}$ will be included in the discussion of the transitional types of functions.

It is useful to note that

$$-2 < \delta < 2.$$

To show this, using a well-known device, we see that

$$\int_r^s f(t)\ (t^2 + \lambda t)^2\, \alpha t = \alpha_4 + 2\lambda\alpha_3 + \lambda^2$$

is never negative since $f(t) \geqq 0$, $r \leqq t \leqq s$, for any real λ. This requires that

$$\alpha_3^2 \leqq \alpha_4 .$$

But

$$-2 + \frac{4\alpha_4 - 3\alpha_3^2}{\alpha_4 + 3} = \delta = 2 - \frac{\alpha_3^2 + 4}{\alpha_4 + 3}$$

and the result follows. One consequence of this is that b_0 cannot vanish for any Pearson frequency function possessing moments of the fourth order.

Turning now to the integration of (1) and the development of the various forms of $f(t)$ that arise, it is useful to make the preliminary statements:

1. Over the range of variation of t, we must have $f(t) \geqq 0$.
2. The area under curve $y = f(t)$ over the range of variation must be finite. This being true then we always determine the constant of integration so that this area is unity.
3. The range in each case is taken as the maximum one for which (1) and (2) may be secured which contains the point, $t = 0$.
4. It is sufficient throughout to take $\alpha_3 \geqq 0$ since the curve for $\alpha_3 = -k$ is only a reflection of that for $\alpha_3 = k$ through the line $t = 0$.

[1] See the Handbook, pp. 103, 104.

It seems best to follow the Handbook in disposing of three of the transitional types before proceeding to the main types of the system and then to the remaining transitional types.

The discussion is planned to embody a direct and uniform method of treatment, giving simple formulae for the calculation of the parameters in terms of α_3 and δ in each case, and noting the salient features of each type of curve. The criteria for each type are expressed in terms of α_3 and δ, which for the whole system permit a simple graphical representation by means of the chart found at the end of this article. The construction of this chart is made clear in the deviation of the criteria.

Transitional Type: The Normal Frequency Function: $\alpha_3 = \delta = 0$

In this case (1) reduces to,

$$\frac{1}{y}\frac{dy}{dt} = -t,$$

from which

$$\text{(N)} \qquad y = c\,e^{-\frac{t^2}{2}}.$$

The range is, of course, $(-\infty, \infty)$ with $C = (2\pi)^{-\frac{1}{2}}$. On the chart, which we shall refer to as the (α_3^2, δ)-diagram, we see that this function corresponds to but a single point.

It may have the appearance of reasoning in a circle to use the values of the parameters given by (3), which were derived from (2), in solving (1) and then for the solution obtained examine the validity of (2). However, we may argue as follows: We will use the relations (3) as definitions of a, b_0, b_1, and b_2 in terms of α_3 and δ which are not yet defined. Using the values of a and the b's given by any choice of α_3 and δ, we solve (1). If the solution is such that for it (2) may be derived, then the relations (3) are valid when α_3 and δ have their usual meanings. For convenience let us denote the conditions for the validity of (2) by (A). It is obvious that conditions (A) are satisfied for

$$\text{(N)} \qquad f(t) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{t^2}{2}}.$$

$$\textbf{Transitional types} \begin{cases} \textbf{III, } \alpha_3 \neq 0, & \delta = 0 \\ \textbf{X, if also } \alpha_3^2 = 4. \end{cases}$$

We get here (See the Handbook, loc. cit.):

$$\text{(III)} \qquad f(t) = \frac{A^{A^2}\,e^{-A^2}}{\Gamma(A^2)}\,(A + t)^{A^2-1}\,e^{-At},$$

if $A = 2/\alpha_3$, the range being $(-A, \infty)$.

It is readily verified that, since $A^2 - 1 > -1$, conditions (A) are satisfied.

For $A^2 > 1$ (i.e., for $\alpha_3^2 < 4$) the curve is bell-shaped; for $A^2 < 1$ it is J-shaped with an infinite ordinate at $t = -A$. For the bell-shaped curve the mode falls at $t = -1/A$ and the mean—the mode $= 1/A = \alpha_3/2$.

For $A^2 = 1$, we have

$$\text{(X)} \qquad f(t) = \frac{e^{-t}}{e},$$

which represents a J-shaped curve with the range $(-1, \infty)$.

For $A^2 \neq 1$, the function has been designated type III, the special case as type X. On the (α_3^2, δ)-chart the points corresponding to type III functions fall on the line $\delta = 0$, the type X functions being represented by a single point on this line.

Turning now to the discussion of the three main types, we note that for $\delta \neq 0$, $b_2 \neq 0$ and that consequently the denominator on the right in (1) is always a quadratic which we can write in the form

$$b_2(t - r_1)(t - r_2)$$

in which neither r_1 nor r_2 can be zero (since $b_0 \neq 0$), and

$$\text{(4)} \qquad r_1 = \frac{-b_1 + \sqrt{b_1^2 - 4b_0b_2}}{2b_2} = \frac{-\alpha_3 - \sqrt{\alpha_3^2 - 4\delta(\delta + 2)}}{2\delta} = \frac{-\alpha_3 + \sqrt{D}}{2\delta}.$$

$$r_2 = \frac{-\alpha_3 - \sqrt{D}}{2\delta}.$$

Leaving aside the special case, $r_1 = r_2$, to be dealt with later, we can always solve (1) in the form

$$\text{(5)} \qquad f(t) = C(t - r_1)^{m_1}(t - r_2)^{m_2}$$

with

$$\text{(6)} \qquad m_1 = \frac{a - r_1}{b_2(r_1 - r_2)} = \frac{1 + \delta}{\delta}\frac{\alpha_3}{\sqrt{D}} - \frac{1 + 2\delta}{\delta}$$

$$m_2 = \frac{a - r_2}{b_2(r_2 - r_1)} = -\frac{1 + \delta}{\delta}\frac{\alpha_3}{\sqrt{D}} - \frac{1 + 2\delta}{\delta}.$$

For $\delta < 0$, the r's are real and opposite in sign; for $\delta > 0$ and $\alpha_3^2 < 4\delta(\delta + 2)$, the r's are complex; and for $\delta > 0$ and $\alpha_3^2 > 4\delta(\delta + 2)$, the r's are real and of the same sign. These three conditions with the additional condition that $\alpha_3 \neq 0$ give rise respectively to the *main* types of frequency functions designated I, IV, and VI. The points corresponding to them fall in simply determined areas on the (α_3^2, δ)-chart. The boundaries of these areas, the curve,

$$(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2(2 + \delta),$$

which intersects the type I and type VI areas, and the line,

$$\delta = -1/2$$

contain the points which correspond to the transitional types.

Main Type I. $\alpha_3 \neq 0, -1 < \delta < 0\ [\delta \neq -\frac{1}{2}, (2+3\delta)\alpha_3^2 \neq 4(1+2\delta)^2(2+\delta)]$

For $\alpha_3 > 0$, we see that

$$r_1 < 0 < r_2 \text{ and that } |r_1| < |r_2|\,.$$

The range is taken to be (r_1, r_2) and (5) is written

$$y = C(t - r_1)^{m_1}(r_2 - t)^{m_2}\,. \tag{I}$$

It is evident that the area under the curve over this interval is finite only when $m_1 + 1 > 0$ and $m_2 + 1 > 0$ and that if these inequalities hold moments of all orders exist. In this case also conditions (A) are satisfied. Now

$$m_1 + 1 = -\frac{1+\delta}{\delta}\left(1 - \frac{\alpha_3}{\sqrt{D}}\right)$$

$$m_2 + 1 = -\frac{1+\delta}{\delta}\left(1 + \frac{\alpha_3}{\sqrt{D}}\right),$$

and in the present case

$$1 \pm \frac{\alpha_3}{\sqrt{D}} > 0\,.$$

Thus $m_1 + 1$ and $m_2 + 1$ are each > 0 only if $\delta > -1$. On the chart, then, the points for $\delta < -1$ correspond to no frequency functions,—they fall in the "Impossible Area."

Further the type I curve will be U-shaped, J-shaped, or bell-shaped if both m's are < 0, if the m's are opposite in sign, or if both are > 0. We have

$$m_1 = -\frac{1+\delta}{\delta}\left(1 - \frac{\alpha_3}{\sqrt{D}}\right) - 1\,.$$

Since for $-1 < \delta < -\frac{1}{2}$,

$$0 < -\frac{1+\delta}{\delta} < 1\,,$$

we see that $m_1 < 0$ $(\alpha_3 > 0)$ for δ in this interval. For $-\frac{1}{2} < \delta < 0$, $m_1 > 0$ only if

$$-\frac{1+\delta}{\delta}\left(1 - \frac{\alpha_3}{\sqrt{D}}\right) > 1\,,$$

which leads to the condition:

$$(2 + 3\delta)\alpha_3^2 < 4(1 + 2\delta)^2 (2 + \delta) .$$

Also,

$$m_2 = -\frac{1 + \delta}{\delta}\left(1 + \frac{\alpha_3}{\sqrt{D}}\right) - 1$$

whence it is similarly seen that $m_2 > 0$ when $-\frac{1}{2} < \delta < 0$, and that generally $m_2 > 0$ only when

$$(2 + 3\delta)\alpha_3^2 < 4(1 + 2\delta)^2 (2 + \delta) .$$

Thus the curve,

$$(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2 (2 + \delta) ,$$

being tangent to the line $\alpha_3^2 = 0$ at $\delta = -\frac{1}{2}$, divides the type I area on the chart into three parts: Above it lie the points corresponding to U-shaped curves, to the right of it the points corresponding to J-shaped curves, and below it the points corresponding to bell-shaped curves. (Note that for $\delta < -\frac{2}{3}$ the curves are always U-shaped.)

Since $r_2 - r_1 > 0$ and $b_2 \gtrless 0$ accordingly as $\delta \lesseqgtr -\frac{1}{2}$, it is readily verified that $r_1 < a < r_2$ only for U- or bell-shaped curves. The sign of a is always opposite to that of α_3 for curves with a mode. Finally the constant is determined by setting

$$C \int_{r_1}^{r_2} (t - r_1)^{m_1}(r_2 - t)^{m_2}\, dt = 1 ,$$

giving

$$C = \frac{1}{\beta(m_1 + 1, m_2 + 1)\,(r_2 - r_1)^{m_1 + m_2 + 1}} .$$

Main Type IV: $\alpha_3 \neq 0$, $\delta > 0$, and $\alpha_3^2 < 4\delta(\delta + 2)$

In this case we write:

$$r_1 = \frac{-\alpha_3}{2\delta} + \frac{i\sqrt{-D}}{2\delta} = -r + is, \qquad r_2 = -r - is .$$

$$m_1 = -\frac{1 + \delta}{\delta}\frac{\alpha_3}{\sqrt{-D}}\, i - \frac{1 + 2\delta}{\delta} = \frac{\nu i}{2} - m, \qquad m_2 = -\frac{\nu i}{2} - m .$$

With this notation (5) becomes

$$y = C[(t + r)^2 + s^2]^{-m}\left(\frac{t + r - is}{t + r + is}\right)^{\frac{\nu i}{2}},$$

and since,

$$\left(\frac{a-bi}{a+bi}\right)^{\frac{ci}{2}} = e^{c\tan^{-1}b/a} = e^{c(\pi/2-\tan^{-1}a/b)},$$

the frequency function can be written,

$$\text{(IV)} \qquad y = Ce^{\nu\pi/2}[(t+r)^2+s^2]^{-m}e^{-\nu\tan^{-1}\frac{t+r}{s}}.$$

It is readily seen that $m > 0$, that ν is opposite in sign to α_3, that

$$e^{-\nu\tan^{-1}\frac{t+r}{s}}$$

can always be taken to lie between $e^{-\nu\pi/2}$ and $e^{\nu\pi/2}$, and that the range can be taken $(-\infty, \infty)$.

In the previously discussed cases in which $\delta \leqq 0$, if the area under the was finite moments of all orders existed. In the present case, the area and first four moments are always finite but this may fail to be true of moments higher orders. For, since $0 < \delta < 2$,

$$m = \frac{1+2\delta}{\delta} > \frac{5}{2},$$

and the integral,

$$\int_{-\infty}^{\infty} t^n f(t)\,dt$$

for $f(t)$ given by (IV) will be finite for $n \leqq 4$ and infinite for $n = 5$ if δ In order for the n-th moment to exist we must have

$$2m > n+1$$

or

$$\delta < \frac{2}{n-3}.$$

Pearson designated as *heterotypic* those members of his system of frequency functions for which the eighth moment failed to exist. (In such a case the standard deviation of the fourth moment in samples would be infinite.) Setting $n = 8$, we get $\delta = 2/5$ as the deadline on the (α_3^2, δ)-chart.

It was apparent that conditions (A) were satisfied for $-1 < \delta < 0$. (It will appear below that the case in which $\delta = -\frac{1}{2}$ is no exception.) For $\delta > 0$ it will be seen that it is generally true, as in the present case, that the formulae (2) and (3) can be derived if α_{n+2} exists, i.e., if

$$\delta < \frac{2}{n-1}.$$

To determine C, on setting the integral of (V) over the interval $(-\infty, \infty)$ equal to unity, we get

$$C = \frac{s^{2m-1}}{G(2m-2, \nu)}$$

in which

$$G(2m-2, \nu) = \int_0^\pi \sin^{2m-2}\varphi e^{\nu\varphi}\, d\varphi \,^{5} \qquad \left(\varphi = \frac{\pi}{2} - \tan^{-1}\frac{t+r}{s}\right).$$

Main Type VI: $\alpha_3 \neq 0, \delta > 0, \alpha_3^2 > 4\delta(\delta+2)\,[(2+3\delta)\alpha_3^2 \neq 4(1+2\delta)^2(2+\delta)]$

The conditions specify the remaining area on the chart. This may be left in the form

$$y = C(t-r_1)^{m_1}(t-r_2)^{m_2}. \tag{5}$$

Now r_1 and r_2 are both opposite in sign to α_3, which, as usual, we will consider positive, and $|r_2| > |r_1|$. Always $m_2 < 0$ and $m_1 \gtrless 0$ accordingly as

$$(2+3\delta)\alpha_3^2 \lesseqgtr 4(1+2\delta)^2(2+\delta).$$

We note that

$$a - r_2 = b_2(r_2 - r_1)m_2 > 0,$$

since now $b_2 > 0$, and that

$$a - r_1 = b_2(r_1 - r_2)\, m_1$$

has the same sign as m_1. Finally $a < 0$.

Thus for $\alpha_3 > 0$ and $m_1 > 0$, the point $t = a$ on the axis of t lies to the right of both $t = r_1$ and $t = r_2$. Also

$$m_1 + m_2 = -\frac{2(1+2\delta)}{\delta} = -4 - \frac{2}{\delta}.$$

The range is taken (r_1, ∞), the curve being bell-shaped when $m_1 > 0$. If $m_1 < 0$, the curve is J-shaped, $t = a$ now lying to the left of $t = r_1$.

Since

$$m_1 + m_2 < -5, \text{ and } m_1 + 1 > 0,$$

the area and the first four moments always exist. In order for the n-th moment to be finite, we must have

$$-(m_1 + m_2) > n + 1$$

which is the same condition as in the case of the type IV function, giving the same deadline, $\delta = 2/5$.

---

[5] Cf: Tables for Statisticians and Biometricians, Cambridge Univ. Press, Part I, 2nd edition (1924), p. lxxxi.

If the origin be shifted to the point, $t = r_2$, we have writing,

$$t - r_2 = z, \qquad r_1 - r_2 = \alpha,$$

for the type VI function the expression,

$$y = Cz^{m_2}(z - \alpha)^{m_1}, \tag{VI}$$

with the range (α, ∞). Finally

$$C = \frac{1}{\alpha^{m_1+m_2+1}\beta(m_1 + 1, -m_1 - m_2 - 1)}.$$

Transitional Type II: $\alpha_3 = 0, -1 < \delta < 0. \quad (\delta \neq -\frac{1}{2})$

In this case,

$$r_1 = -r_2 = \frac{\sqrt{D}}{\delta} < 0$$

$$m_1 = m_2 = -\frac{1 + 2\delta}{\delta} \gtrless 0 \quad \text{accordingly as } \delta \gtrless -\tfrac{1}{2}.$$

The frequency function is a special case of type I; setting,

$$-r_1 = r_2 = S$$

$$m_1 = m_2 = M,$$

we can write it in the form,

$$y = C(S^2 - t^2)^M. \tag{II}$$

As in all cases in which $\alpha_3 = 0$, the curve is symmetrical about the mean.[6] As in the type I case, the area and moments do not exist for $\delta \leqq -1$; for $-1 < \delta < -\frac{1}{2}$, the curve is U-shaped; for $-\frac{1}{2} < \delta < 0$, it is bell-shaped. The range is, of course, $(-S, S)$.

Finally,

$$C = \frac{1}{(2S)^{2M+1}\beta(M + 1, M + 1)}.$$

Transitional Type VII; $\alpha_3 = 0, \delta > 0$

This function may be regarded as a special case of type IV, with

$$r = 0, \quad s = \frac{\sqrt{4\delta(\delta + 2)}}{2\delta} > 0, \quad \nu = 0, \quad \text{and} \quad m = \frac{1 + 2\delta}{\delta} > 0,$$

[6] It follows at once from the recursion formula,

$$\alpha_{n+1} = \frac{n}{2 - (n - 2)\delta}\,[(2 + \delta)\,\alpha_{n-1} + \alpha_3\,\alpha_n],$$

obtained from setting the expressions (3) in (2), that on changing the sign of α_3, the signs of all the odd moments are changed.

and we write the function:

(VII) $$y = C(t^2 + s^2)^{-m}.$$

The type VII function may equally well be derived from the type II function by noting that

$$S = is \text{ and } M = -m.$$

The range is $(-\infty, \infty)$ however and for $\delta \geqq 2/5$ the function is heterotypic. Finally

$$C = \frac{s^{2m-1}}{\sqrt{2\pi}} \frac{\Gamma(m)}{\Gamma\left(\frac{2m-1}{2}\right)}.$$

Transitional Type V; $\alpha_3 \neq 0$, $\delta > 0$, $\alpha_3^2 = 4\delta(\delta + 2)$

Here

$$r_1 = r_2 = -r$$

and we return to (1) to derive the form of the function, writing it: (The type V can also be derived as a limiting form of type VI)

$$\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{b_2(t + r)^2}.$$

On integration we get

$$y = C(t + r)^{-\frac{1}{b_2}} e^{-\frac{a+r}{b_2(t+r)}}$$

$$= C(t + r)^{-\frac{2(1+2\delta)}{\delta}} e^{-\frac{\alpha_3(1+\delta)}{\delta^2(t+r)}}.$$

(V) $$= C(t + r)^{-2m} e^{-\frac{2r(m-1)}{t+r}}.$$

We note that r has the same sign as α_3 and that $m = 2 + 1/\delta$. The range is taken to be $(-r, \pm \infty)$ accordingly as $\alpha_3 \gtrless 0$. The curve is always bell-shaped. In order for the n-th moment to exist we must have as always when $\delta > 0$,

$$4 + 2/\delta > n + 1$$

leading to the same conclusions as in the type IV or VI case. Finally

$$C = \frac{[2r(m-1)]^{2m-1}}{\Gamma(2m-1)}.$$

Transitional Type VIII; $\alpha_3 = 0$, $\delta < -\frac{1}{2}$, $(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2(2 + \delta)$

The function is a special case of type I in which $m_1 < 0$ and $m_2 = 0$. But when $m_2 = 0$, $m_1 = -2m$, and the frequency function becomes

(VIII) $$y = C(t - r_1)^{-2m}.$$

The range is (r_1, r_2), the curve being J-shaped with an infinite ordinate at $t = r_1$ and a finite one at $t = r_2$. In this case,

$$C = \frac{1 - 2m}{(r_2 - r_1)^{1-2m}}. \qquad (1 - 2m > 1)$$

Transitional Type IX: $\alpha_3 \neq 0$, $-\frac{1}{2} < \delta < 0$, $(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2(2 + \delta)$

We have another special type I function in which $m_1 = 0$ and $m_2 = -2m > 0$. The function is

$$y = C(r_2 - t)^{-2m} \qquad \text{(IX)}$$

the range still being (r_1, r_2), the curve being J-shaped with a finite ordinate at $t = r_2$. C has the same value as in the type VIII case.

Transitional Type XI; $\alpha_3 \neq 0$, $0 < \delta < 2/5$, $(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2(2 + \delta)$

The function is a special type VI in which $m_1 = 0$, and $m_2 = -2m < 0$, and we may write it

$$y = C(t - r_2)^{-2m} \qquad \text{(XI)}$$

with the range still (r_1, ∞). The curve is J-shaped with a finite ordinate at $t = r_1$. Again,

$$C = \frac{2m - 1}{(r_1 - r_2)^{2m-1}} \qquad \left(2m - 1 = 3 + \frac{2}{\delta}\right)$$

Transitional Type XII: $\delta = -\frac{1}{2}$

If $\delta = -\frac{1}{2}$, the four linear equations derived from (2) from which the values of a, b_0, b_1, and b_2 in (3) are derived are inconsistent. We can however set the values (3) in the differential equation (1) and from its limiting form as $\delta \to -\frac{1}{2}$, derive the function appropriate to this case.

We obtain

$$\frac{1}{y}\frac{dy}{dt} = \frac{-\alpha_3 - 2(1 + 2\delta)t}{(2 + \delta) + \alpha_3 t + \delta t^2}$$

and if $\delta = -\frac{1}{2}$, this becomes

$$\frac{1}{y}\frac{dy}{dt} = \frac{2\alpha_3}{t^2 - 2\alpha_3 t - 3} = \frac{2\alpha_3}{(t - r_1)(t - r_2)}$$

with

$$r_1 = \alpha_3 - \sqrt{\alpha_3^2 + 3}, \qquad r_2 = \alpha_3 + \sqrt{\alpha_3^2 + 3}.$$

On integration,

$$y = C'(t - r_1)^{m_1}(t - r_2)^{m_2},$$

in which

$$m_1 = -\frac{\alpha_3}{\sqrt{\alpha_3^2 + 3}}, \qquad m_2 = \frac{\alpha_3}{\sqrt{\alpha_3^2 + 3}}.$$

We observe that ($\alpha_3 > 0$)

$$r_2 > 0 > r_1, \qquad |r_2| > |r_1|$$

$$m_2 = -m_1 > 0.$$

Taking the range to be (r_1, r_2), we write,

(XII) $$y = C\left(\frac{r_2 - t}{t - r_1}\right)^{m_2},$$

the curve being J-shaped. Here

$$C = \frac{1}{(r_2 - r_1)\,\beta(1 - m_2,\, 1 + m_2)}.$$

The values of the parameters and the form of the function can also be derived as a special type I function in which $\delta = -\frac{1}{2}$.

Finally we note that for $\alpha_3 = 0$, (XII) reduces to

$$y = C$$

thus including the rectangular distribution function among the Pearson system.

In the course of the above discussion a system of criteria for the various types of functions has been set up in terms of α_3 and δ, in terms of which in every case the parameters may be readily calculated. The (α_3^2, δ)-chart which makes these criteria visual is comparatively simple to construct and is strikingly simple in appearance. Besides the lines,

$$\delta = -1, \qquad \delta = -\frac{1}{2}, \qquad \delta = 0, \qquad \delta = \frac{2}{5}, \quad \text{and} \quad \alpha_3 = 0,$$

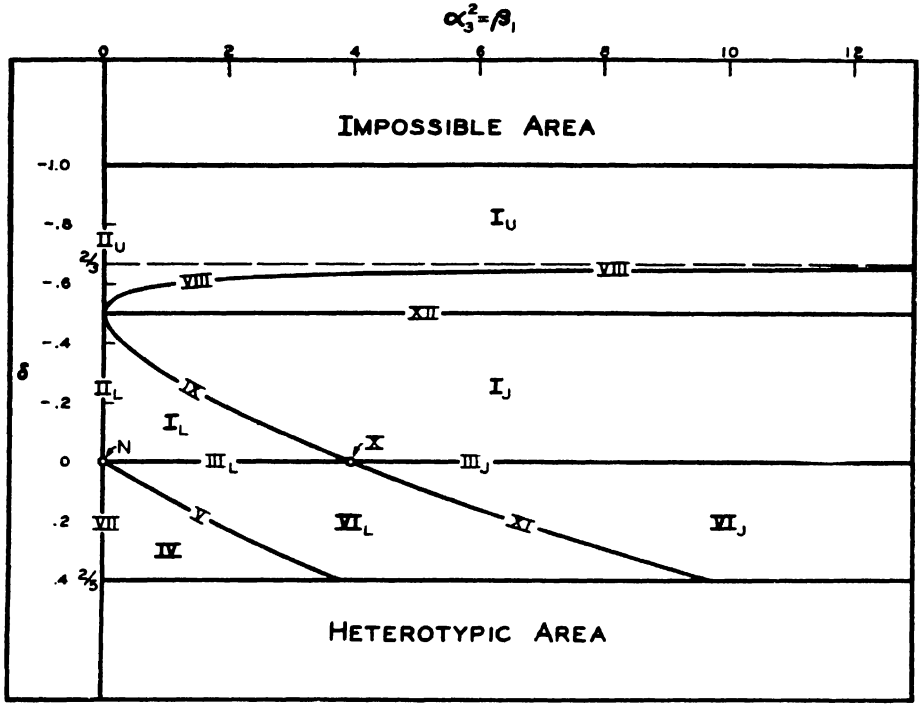
it contains only the curves

$$\alpha_3^2 = 4\delta(\delta + 2)$$

on which the points corresponding to the type V function lie, and the curve,

$$(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2\,(2 + \delta)$$

on which the points corresponding to the functions of types VIII, IX, X, and XI are found. I must take occasion to express my thanks to Mr. Simon Yang who constructed this chart for me.



THE (α^2, δ) CHART FOR THE PEARSON SYSTEM OF FREQUENCY CURVES
(The subscript L refers to bell-shaped curves)

THE UNIVERSITY OF MICHIGAN.

RANK CORRELATION AND TESTS OF SIGNIFICANCE INVOLVING NO ASSUMPTION OF NORMALITY*,†

By Harold Hotelling and Margaret Richards Pabst

1. Dependence of Tests of Significance on Normality

The powerful tests of significance, largely the work of R. A. Fisher, which have been revolutionizing statistical theory and practice, are in the main based on the assumption of a normal distribution in a hypothetical population from which the observations are a random sample. The nature and extent of the errors likely to result from the application of a test of significance assuming normality, where normality does not really exist, have been the subject of investigations both experimental and mathematical,[1] which however have not produced satisfactory substitutes for Fisher's methods. A false assumption of normality does not usually give rise to serious errors in the interpretation of simple means, since the distribution of a mean of any considerable number of cases is very nearly normal, no matter what the nature of the parent population, so long as it does not fall within a certain class having infinite range, and including the Cauchy distribution. The sampling distributions of second-order statistics are however more seriously disturbed by lack of normality, as is evident from their standard errors. For example the variance $(\mu_4 - \mu_2^2)/n$ of sample variances is much affected if μ_4/μ_2^2 differs considerably, as it often does, from the value 3 which it takes for a normal distribution. Likewise the approximate variance of the correlation coefficient,

$$\sigma_r^2 = \frac{1}{n\mu_{20}\mu_{02}}\left\{\mu_{22} + \frac{\mu_{40}\mu_{11}^2}{4\mu_{20}^2} + \frac{\mu_{04}\mu_{11}^2}{4\mu_{02}^2} - \frac{\mu_{31}\mu_{11}}{\mu_{20}} - \frac{\mu_{13}\mu_{11}}{\mu_{02}} + \frac{\mu_{22}\mu_{11}^2}{2\mu_{20}\mu_{02}}\right\},$$

* Research under a grant-in-aid from the Carnegie Corporation of New York.

† Presented to the American Mathematical Society at New York, Oct. 26, 1935.

[1] J. L. Carlson, *A Study of the Distribution of Means Estimated from Small Samples by the Method of Maximum Likelihood for Pearson's Type II Curve*, Unpublished M. A. Thesis, Leland Stanford Junior University, 1931.

Leone Chesire, Elena Oldis and Egon S. Pearson, *Further Experiments on the Sampling Distribution of the Correlation Coefficient*, Journal of the American Statistical Association, June, 1932, pp. 121–128.

Victor Perlo, *On the Distribution of Student's Ratio for Samples of Three Drawn from a Rectangular Distribution*, Biometrika, Vol. XXV, Parts I and II, May, 1933, pp. 203–204.

Paul R. Rider, *On the Distribution of the Ratio of Mean to Standard Deviation in Small Samples from Non-Normal Universes*, Biometrika, Vol. XXI, Parts I to IV, December, 1929, pp. 124–143.

H. L. Rietz, *Note on the Distribution of the Standard Deviation, etc.*, Biometrika, Vol. XXIII, 1931, pp. 424–426.

W. A. Shewhart and F. W. Winters, *Small Samples—New Experimental Results*, Bell Telephone Laboratories, Reprint B-327, July, 1928.

where μ_{ij} is the mean value of $x^i y^j$, and $\mu_{10} = \mu_{01} = 0$, may be substantially different from the value $(1 - \rho^2)^2/n$ commonly used, to which it reduces if the population has the bivariate normal distribution. It is however remarkable that if the variates are really independent, so that $\mu_{11} = 0$ and $\mu_{22} = \mu_{20}\mu_{02}$, this formula reduces to

$$\sigma_r^2 = \frac{1}{n}, \tag{1}$$

regardless of the form of the distribution. It should of course be remembered that these formulae give only the first term of an expansion in inverse powers of n, and also that the standard error fails for small samples to characterize the distribution adequately. But the sensitiveness of the standard error formula to deviations from normality in the population is a symptom of the grave dangers in using even those distributions which for normal populations are accurate, in the absence of definite evidence of normality.

To substitute in standard error formulae values of the higher moments estimated from the data does not meet the difficulty satisfactorily, since these higher moments are themselves subject to sampling errors which are often large, and since no exact distributions can ever be obtained in this way. The use of an arbitrary system of distributions such as the Pearson curves is subject to the same criticisms as that of the normal distribution. These and other special distributions may indeed be justified in special cases by general reasoning; an example of this in introducing a measure of relationship other than the correlation coefficient is to be found in the genetic discussion of Chapter 9 of Fisher's "Statistical Methods for Research Workers." But for a great deal of statistical work no such a priori reasoning is available and sufficient to specify a distribution in sufficient detail. If a specific form of distribution other than the normal can be relied on in a particular case, the mathematical problem of finding the exact distribution of the appropriate statistic will still commonly be found difficult or impossible.

2. Tests Independent of Normality Assumptions

A set of problems is thus encountered regarding the nature and methods of statistical inference possible without assuming any particular distribution of the variates in the population from which we have a sample. Tests of significance underlying such inferences must clearly be invariant under all transformations of each variate. We are thus forced to rely for our information on relations of *order*, or of qualitative classification, rather than upon magnitudes, excepting insofar as we can use inequalities such as that of Tchebycheff. Classification leads to the use of contingency tables, from which accurate probabilities are calculable for testing whether or not the two or more principles of cross-classification used are independent. If the probability obtained is so small as to render it incredible that independence exists, the further problem arises of measuring the degree of relationship; but in the absence of special assumptions, such as that

of the bivariate normal distribution, or those in Fisher's genetic example mentioned above, the problem of measuring degree of relationship is insoluble. Any measure of degree of relationship will change its value, unless this value corresponds to independence, when transformations other than those of a restricted class are applied to one of the variates. The problem of measuring *degree* of relationship, or correlation, is thus of quite a different character from that of testing the *existence* of a relationship, which is equivalent to absence of independence. The existence of correlation may be detected by methods of rank order or of classification; these can never, by themselves, be sufficient for its measurement.

To test the deviation of the center of a symmetrical population from some definite hypothetical value, Student's distribution, which is appropriate when the population is normal, may be replaced by the binomial distribution, which will sometimes show that the preponderance of cases on one side of the hypothetical value is too great to admit the hypothesis. Fisher applied this principle to Student's original example, showing at the same time that it can in certain cases be used to test the significance of the difference between the means of two samples.[2] Both this type of test and the use of contingency tables with grouped values of variates bring out clearly the fact that abandonment of the assumption of normality is equivalent to a certain loss of information, larger samples being required to make up for the lack of knowledge of the form of the population. The loss of information is greater for contingency tables arranged according to the values of the variates than when an appropriate method of rank correlation is used, for the contingency table may be regarded as derived from the ranks by grouping them, thus discarding some of the information.

We shall in §8 illustrate a combination of rank and contingency methods suitable for utilizing simultaneously two kinds of information contained in grouped data.

For large samples a method of treatment for which a great deal is to be said in many cases consists of replacing the observed variate by a new variate x to which a value is assigned for each individual or frequency class by interpolation in a table of the normal probability integral, in such a way that the distribution of x in the sample approximates normality. If this is done for each of two variates which do not have the bivariate normal distribution, the transformed values x and y may also lack the bivariate normal distribution, even approximately, though each is normally distributed, so far as we can speak of a sample as being normally distributed. Even if the bivariate distribution is normal, the correlation coefficient of x and y will not have the same distribution as the correlation coefficient in samples drawn from a bivariate normal distribution, since in the latter case the distributions of x and y separately would in most samples be less nearly normal than when the transformation to approximate normality is applied. From these considerations it follows that for the detection of correlation the normalizing transformation cannot be said in general to be the best

[1] R. A. Fisher, *Statistical Methods for Research Workers*, Art. 24, end.

method, even for large samples, though it may be a useful preliminary to the application of the method of least squares or to the use of correlation coefficients significantly different from zero in certain cases.

3. The Rank Correlation Coefficient

Suppose that n individuals are arranged in two orders with respect to two different attributes. Thus we might arrange a freshman class in order according to their grades in a language examination, and also according to their mathematical grades. As another example, we might be able to obtain ratings of various states with respect to penal law or practice, and also with respect to amount of crime. Continuous variates expressing these qualities are likely not to be normally distributed, so that the product-moment correlation coefficient r cannot be expected to have the exact distribution known for it in the case of samples from a normal population. We may therefore resort to the ranks, ignoring any exact values that have been assigned.

Calling X_i the rank of the ith individual with respect to one attribute, and Y_i his rank with respect to the other, so that $(X_1, X_2, \cdots, X_n)$ and $(Y_1, Y_2, \cdots, Y_n)$ are two permutations of the numbers $(1, 2, \cdots, n)$, let us put $x_i = X_i - \bar{x}$, $y_i = Y_i - \bar{y}$, where

$$\bar{x} = \bar{y} = \frac{n+1}{2}$$

The rank correlation coefficient is defined as

$$r' = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}}, \tag{2}$$

the sums being over the n values in the sample. Now since the sum of the first n integers is $n(n+1)/2$, and the sum of their squares is $n(n+1)(2n+1)/6$, we have

$$\begin{aligned} \Sigma x^2 = \Sigma (X - \bar{x})^2 &= \Sigma X^2 - (\Sigma X)^2/n \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n^3 - n}{12}, \end{aligned} \tag{3}$$

and Σy^2 has the same value. Also, if we put d_i for the difference between the two ranks for the ith individual, so that

$$d_i = X_i - Y_i = x_i - y_i,$$

we have

$$\Sigma d^2 = \Sigma x^2 - 2\Sigma xy + \Sigma y^2 = \frac{n^3 - n}{6} - 2\Sigma xy.$$

Substituting in (2) the value of Σxy found from this equation, and also the values just obtained for Σx^2 and Σy^2, we have:

$$r' = 1 - \frac{6\Sigma d^2}{n^3 - n}. \tag{4}$$

This is the most convenient formula for computing r'.

Compared with certain other tests of correlation based on order, such as $\Sigma|d|$, or the number of inversions required to pass from one permutation of the n numbers to the other, r' appears to be a sensitive index of relationship, since for a given value of n it possesses a greater number of distinct values. But to assert without qualification that r' or any other statistic is the best possible test of correlation based on order relations alone would be meaningless. Indeed, a particular type of bivariate distribution might well have a parameter representing correlation whose significance could best be detected by a test adapted only to this particular bivariate distribution. However the rank correlation coefficient has properties that point to its value in more general use than it has heretofore received. It has been regarded chiefly as a more easily calculable substitute for the product-moment coefficient r. Karl Pearson has remarked that the rank correlation coefficient is the easier to compute for samples smaller than approximately forty, while r involves less labor for larger samples.

The great value of the rank correlation coefficient appears to us to consist in its use as a test of the existence of correlation, a test capable of exact interpretation in terms of probability, without any assumption of a normal or other special bivariate distribution. If a bivariate distribution is specified by $f(x, y)\,dx\,dy$, the condition of independence is that $f(x, y)$ shall be the product of a function of x by a function of y. If we put

$$\xi = \int_{-\infty}^{\infty}\int_{0}^{x} f(x', y')\,dx'\,dy', \qquad \eta = \int_{0}^{y}\int_{-\infty}^{\infty} f(x', y')\,dx'\,dy', \tag{5}$$

using the inner integral sign in each case to correspond to the inner differential, then each of the quantities ξ and η is distributed with uniform density from $-\frac{1}{2}$ to $+\frac{1}{2}$; and if x and y are independent, then ξ and η are also independent. The correlation ρ' of ξ with η may be called the rank correlation of x and y in the population. It will vanish in case of independence. It is for this case that we shall obtain in §§5, 6 and 7 the exact probability test for r' in small samples, the exact standard error and fourth moment, and asymptotic values for the higher moments, with a demonstration that, for sufficiently large samples, r' can be treated as normally distributed. In §9 we shall present, in a revised and simplified form, certain work of Karl Pearson relative to the estimation of the correlation ρ in a bivariate normal distribution, and apply the results to discuss the question of the importance of the lost information when measurements are replaced by ranks.

4. History of Rank Correlation Theory

Rank correlation seems to have had its origin in the method of representing the distribution of a variate by grades or percentiles introduced by Francis

Galton.[3] Later Spearman[4] proposed that rank be considered in place of the variate, and suggested that the correlation of ranks be used as a measure of the degree of dependence of the variates. Spearman also introduced the "footrule of correlation" based on $\Sigma | d |$.

The principal memoir on rank correlation is by Karl Pearson.[5] Assuming an underlying normal distribution, Pearson obtains a relation equivalent to

$$\rho = 2 \sin \frac{\pi}{6} \rho', \tag{6}$$

where ρ is the correlation of x and y in the population, and ρ' is the correlation of uniformized variates ξ and η defined by (4). An estimate r'' of ρ may be based on the rank correlation r', in accordance with (6), by writing

$$r'' = 2 \sin \frac{\pi}{6} r'. \tag{7}$$

Pearson finds the first few terms of infinite series giving the standard errors of r' and r''. He deals similarly with the estimation of correlation by means of $\Sigma | d |$. The paper contains a neat proof, attributed to Student, of the probable error of r' under conditions of independence. It was this proof that suggested the analysis of §§6 and 7 below. This long memoir is very difficult to read and interpret accurately, owing chiefly to the failure to distinguish clearly between sample and population.

The use of the probable error formulae is valid only if the distributions of r' and r'' are sensibly normal. The question of approximate normality thus raised is investigated for the first time in the present paper. In order to use these formulae it is necessary to assume not only (1) that the underlying population has the bivariate normal distribution (an assumption which requires more than that each variate be normally distributed), (2) that the first few terms of the infinite series are enough, and (3) that the distributions of r' and r'' are practically normal, but also (4) that sample values can be put for population values in the formulae, or that population values are known independently or can be assumed. It is probably this last condition that has been least understood and has led to the greatest number of false conclusions regarding the significance of data.

A note by W. C. Eells[6] presents a compilation of numerous textbook versions of the probable errors of r' and r'', all differing from each other and from Pear-

[3] Francis Galton, *Natural Inheritance*, Macmillan, 1889, Chaps. 4 and 5.

[4] C. Spearman, *The Proof and Measurement of Association Between Two Things*, American Journal of Psychology, Vol. 15, 1904.

[5] Karl Pearson, *On Further Methods of Determining Correlation*, Drapers' Company Research Memoirs, Biometric Series IV, Mathematical Contributions to the Theory of Evolution, XVI, London, Dulau, 1907.

[6] W. C. Eells, *Formulas for Probable Errors of Coefficients of Correlation*, Journal of the American Statistical Association, Vol. 24, 1929, p. 170.

son's. Taking Pearson's formulae as correct, without discussing the assumptions implicit in their use, Eells presents a table for calculating the probable errors of r, r' and r''.

5. Significance of Rank Correlation in Small Samples

If the variates are independent we may without loss of generality assign the values $1, 2, \cdots, n$ in order to $X_1, X_2, \cdots, X_n$, and regard the Y's as made up by any one of the $n!$ permutations of these numbers, all permutations being equally probable. The probability of any particular value of r' is thus proportional to the number of permutations giving rise to this value. These may be enumerated with the help of (4). Thus for $n = 2$, each of the values ± 1 has the probability $\frac{1}{2}$. For $n = 3$, the possible values of r' are $-1, -\frac{1}{2}, \frac{1}{2}, 1$, with respective probabilities 1/6, 1/3, 1/3, 1/6. For $n = 4$ the values 1, 4/5, 3/5, 2/5, 1/5, 0 have the respective probabilities 1/24, 1/8, 1/24, 1/6, 1/12, 1/12.

From (2) it is evident that the distribution of r' in case of independence is symmetrical, since each permutation is exactly as probable as that of directly opposite order, and since a change of sign of all the x's or y's changes the sign of r' without affecting its absolute value. It is clear also that the values $r' = \pm 1$, corresponding to the two variates being in the same or opposite orders, are the extreme ones, and have each a probability $1/n!$. The next greatest value of $|r'|$ corresponds to the interchange of two consecutive individuals, who may be selected in $n - 1$ ways and makes $\Sigma d^2 = 2$. Thus the values $\pm(1 - 12/[n^3 - n])$ occur with probability $(n - 1)/n!$ each. Next to these, corresponding to $\Sigma d^2 = 4$, are the values $\pm(1 - 24/[n^3 - n])$, whose probabilities are each $(n - 2)(n - 3)/2(n!)$, since the numbers of pairs of mutually exclusive consecutive pairs in a sequence of n is $(n - 2)(n - 3)/2$. In like manner, but with greater complexity, it appears that the probability of the value $1 - 36/[n^3 - n]$ is $\dfrac{(n - 3)(n - 4)(n - 5) + 12(n - 2)}{6(n!)}$. Easy calculation from these results shows that, if we require for significance a probability $P = .01$ of a value of $|r'|$ as great as or greater than the value observed, then for samples of 5 it is impossible to obtain a significant value; for $n = 6$, significance requires that $r' = \pm 1$; and for $n = 7$ the significant values of $|r'|$ are 25/28 and more. For the less stringent standard $P = .05$, a unit correlation only is significant in a sample of 5; while 29/35 is not, but 31/35 is, significant in a sample of 6.

6. The Standard Error and Fourth Moment

For large samples the exact calculation of probabilities becomes very laborious, and we are forced to resort to approximations. The first step in the available approximations is the determination of the standard deviation of the distribution. The square of this quantity, the second moment or variance of r', may, since the mean value of r' in case of independence is zero, be written

$$\sigma_{r'}^2 = \mu_2 = Er'^2,$$

the symbol E denoting the expectation or mean value of the quantity following. The operation E has the properties that the expectation of a sum is the sum of the expectations of the terms, the expectation of the product of *independent* variates is the product of their expectations, and the expectation of the product of a constant by a variate is the product of the constant by the expectation of the variate. It is particularly to be noted that the first of these properties holds whether the terms of the sum are mutually independent or not.

From (2) and (3) we have

$$r' = \frac{12 \Sigma xy}{n^3 - n}. \tag{8}$$

Now we may regard $x_1, x_2, \cdots, x_n$ as taking the same values in all samples, these values being centered at zero and differing consecutively by unity. The y's are then variates, not independent of each other, taking this same set of values, but in a manner varying from sample to sample by chance. For any particular y, for example that associated with x_1, the chance distribution has moments of the form

$$Ey^p = \frac{\Sigma x^p}{n} = \frac{\Sigma y^p}{n} = \frac{s_p}{n}, \tag{9}$$

if we denote by s_p the sum of the pth powers of the n numbers differing consecutively by unity and centered at zero. It is clear that, for every odd value of p, $s_p = 0$. Also, from (3),

$$s_2 = \frac{n^3 - n}{12}.$$

In view of these facts, we have from (8),

$$\sigma_{r'}^2 = Er'^2 = \frac{E(\Sigma xy)^2}{s_2^2} = \frac{\Sigma x^2 Ey^2 + 2 \Sigma x_1 x_2 E y_1 y_2}{s_2^2},$$

where $\Sigma x_1 x_2$ stands for the sum of all the $n(n-1)/2$ *different* terms obtained by permuting the subscripts. We have

$$E y_1 y_2 = \frac{2 \Sigma x_1 x_2}{n(n-1)};$$

also

$$2 \Sigma x_1 x_2 = s_1^2 - s_2 = -s_2.$$

Combining these results we have:

$$\sigma_{r'}^2 = \frac{1}{s_2^2}\left\{\frac{s_2^2}{n} + \frac{s_2^2}{n(n-1)}\right\} = \frac{1}{n-1}. \tag{10}$$

This is the formula obtained by Student and incorporated in Pearson's memoir.

Any desired moment of r' may be obtained in this manner. However the complexity of the calculation increases rapidly with the order of the moment, and the derivation of even the fourth moment is too long to be included in this paper. The value obtained for the fourth moment is

$$\mu_4 = \frac{3(25n^4 - 13n^3 - 73n^2 + 37n + 72)}{25n(n+1)^2(n-1)^3}.$$

It will be observed immediately that the kurtosis, $\beta_2 = \mu_4/\mu_2^2$, approaches the normal value 3 as n increases.

For values of n which are not small enough for the exact probabilities to be computed easily, the Tchebycheff inequality,

$$P \leqq \frac{1}{(n-1)r'^2}, \tag{11}$$

where P is the probability of a deviation exceeding r', will often be of service. Thus, if $n = 25$ and $r' = .9$, (11) shows that P is less than .05, so that the evidence for existence of a relationship should by an ordinary standard be regarded as significant. However this does not in general give an accurate approximation to P, nor do the similar inequalities involving the higher moments.

7. The Higher Moments and the Approach to Normality

A general moment of r' of even order is defined by

$$\mu_{2\alpha} = E\, r'^{2\alpha} = \frac{1}{s_2^{2\alpha}} E\,(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^{2\alpha}. \tag{12}$$

When the parenthesis is expanded we may take the expectation term by term, regarding the x's as constants. Now

$$E\, y_1^{2\alpha} = \frac{\Sigma\, x_1^{2\alpha}}{n}, \qquad E\, y_1^{2\alpha-1} y_2 = \frac{\Sigma\, x_1^{2\alpha-1} x_2}{n(n-1)},$$

and so forth, the sums on the right in the numerators being symmetric functions of the constants x, taken over all different terms obtained from that written by permuting subscripts, and the denominator being in each case the number of terms in the numerator. Thus

$$\mu_{2\alpha} = \frac{1}{s_2^{2\alpha}}\left\{\frac{(\Sigma\, x_1^{2\alpha})^2}{n} + A\,\frac{(\Sigma\, x_1^{2\alpha-1} x_2)^2}{n(n-1)} + B\,\frac{(\Sigma\, x_1^{2\alpha-2} x_2x_3)^2}{n(n-1)(n-2)} + \cdots\right\}, \tag{13}$$

where the coefficients $A, B, \cdots$ depend on α but not on n. With a view to determining the leading term in the expansion of $\mu_{2\alpha}$ in powers of n^{-1}, we shall select the term in the curly brackets in (13) of highest degree, meaning by the degree of one of these rational fractions the excess of the degree of the numerator over that of the denominator.

The symmetric functions are well known to be expressible as polynomials in

the power-sums s_p. In each term of such a polynomial corresponding to one of our symmetric function of degree 2α, the sum of the subscripts of the s_p's must be 2α, since if all the x's are multiplied by a constant such a polynomial must be multiplied by the 2αth power of the constant. Now s_p is a polynomial of degree $p+1$ in n, if n is even, but vanishes identically if n is odd. Consequently the degree in n of any of the terms of the polynomial in the power-sums must exceed 2α by the number of power-sums appearing in this term. Therefore, the term of highest degree in n obtained, when one of the symmetric functions is expressed in terms of the s_p's and thence in terms of n, must contain the greatest possible number of the s_p's. If p is the number of distinct x's in a term of one of our symmetric functions, this function may be written in the form

$$\begin{aligned}\Sigma x_1^{a_1}x_2^{a_2}\cdots x_p^{a_p} &= c_0 s_{a_1}s_{a_2}\cdots s_{a_{p-1}}s_{a_p} - c_1 s_{a_1+a_p}s_{a_2}\cdots s_{a_{p-1}} \\ &\quad - c_2 s_{a_1}s_{a_2+a_p}\cdots s_{a_{p-1}} - \cdots - c_{p-1}s_{a_1}s_{a_2}\cdots s_{a_{p-1}+a_p} \\ &\quad - c' s_{a_1+a_2+a_p}s_{a_3}\cdots s_{a_{p-1}} - \cdots ,\end{aligned} \tag{14}$$

where $a_1 + a_2 + \cdots + a_p = 2\alpha$, and the c's do not involve n. In the right-hand member of the equation above, the first term involves p of the power-sums, while the remaining terms involve fewer of them. Hence, if all the indices $a_1, a_2, \cdots, a_p$ are even, the first term is a polynomial of degree $2\alpha + p$ in n, while the remaining terms are polynomials of lower degree, and are therefore negligible in comparison with the first term when n is sufficiently large. But if any of the indices a_i are odd, the first term vanishes identically, and the degree of (14), regarded as a polynomial in n, is then less than $2\alpha + p$. Since the sum of the indices is 2α, the number of odd ones among them must be even; let this number be denoted by $2q$, and let the number of even indices be m. Then $p = m + 2q$. The terms of highest degree in the right-hand member of (14) must be obtained by grouping the odd indices in pairs to form the subscripts of the s's. The degree is therefore $2\alpha + m + q$.

In (13), the degree of the denominator of each term in the curly brackets is the number of distinct x's appearing in a term of the symmetric function in the numerator, namely p, or $m + 2q$. Hence the excess of the degree of the numerator over that of the denominator is

$$2(2\alpha + m + q) - (m + 2q) = 4\alpha + m .$$

This will be a maximum when m is a maximum, and is independent of q. The maximum value of m is α, and occurs only for the symmetric function

$$\Sigma x_1^2 x_2^2 \cdots x_\alpha^2 . \tag{15}$$

The term involving this function is therefore the only one in the right-hand member of (13) we need consider. Since this symmetric function contains $n(n-1)(n-2)\cdots(n-\alpha+1)/(\alpha!)$ terms, and since in the expansion of

$$(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^{2\alpha}$$

the coefficient of $x_1^2x_2^2 \cdots x_\alpha^2y_1^2y_2^2 \cdots y_\alpha^2$ is, by the multinomial theorem $(2\alpha)!/2^\alpha$, we have from (13),

$$\mu_{2\alpha} \sim \frac{1}{s_2^{2\alpha}} \frac{(2\alpha)!}{2^\alpha} \frac{\alpha!\,(\Sigma\, x_1^2x_2^2 \cdots x_\alpha^2)^2}{n^\alpha}.$$

To evaluate the symmetric function (15), so far as the term of highest order in, n is concerned, we of course need only the first term of (14), which reduces in this case to

$$\Sigma\, x_1^2x_2^2 \cdots x_\alpha^2 = c_0 s_2^\alpha - \cdots.$$

In the expansion of $s_2^\alpha = (x_1^2 + x_2^2 + \cdots + x_n^2)^\alpha$, the coefficient of (15) is $\alpha!$, which is therefore the reciprocal of c_0. Thus we obtain

$$\mu_{2\alpha} = \frac{(2\alpha)!}{\alpha!\,2^\alpha}\left[\frac{1}{n^\alpha} + \cdots\right],$$

the terms dropped being of higher order in n^{-1}.

The 2αth moment of the quotient of r' by its standard error, that is, of $r'\sqrt{n-1}$, is $(n-1)^\alpha$ times that of r', and therefore approaches, as n increases, the value

$$\frac{(2\alpha)!}{\alpha!\,2^\alpha}. \tag{16}$$

The odd moments are all zero because of the symmetry of the distribution of r'. But (16) is the moment of order 2α of a normal distribution of unit variance and zero mean. It follows therefore from the Second Limit Theorem of Probability[7] that the distribution tends to normality as n increases; that is, for any real number λ, the limit as n tends to infinity of the probability that $r'\sqrt{n-1} < \lambda$ is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-\frac{1}{2}x^2}\, dx.$$

The normality of the limiting distribution of the rank correlation coefficient is rather remarkable, since r', unlike the product-moment correlation coefficient r and other statistics in common use, is neither a mean of independent quantities nor a function of such means, so that the ultimate normality just established is not a corollary of known general theorems. It is unexpected also because the exact distribution of r' for samples smaller than six might lead one to anticipate a bimodal distribution.

An outstanding problem is to determine whether the distribution of r' in samples from a bivariate normal distribution for which $\rho \neq 0$ converges to normality. Without such an approach to normality, the probable error formulae

[7] First proved by Markoff. Cf. Fréchet and Shohat, *A Proof of the Generalized Second Limit Theorem in the Theory of Probability*, Transactions of the American Mathematical Society, Vol. 33, 1932, pp. 533-543.

discovered by Pearson are useless. Another problem is to find convenient and accurate approximations to the distribution of r', for moderate values of n, with close limits of error. A table calculated along the lines suggested in §5 would be very useful.

8. Combination of Rank and Contingency Methods

Suppose that a thousand school children are examined at the end of a course of instruction, and rated with the grades A, B, C and D. Five hundred of these children are of each sex. The results are:

	A	B	C	D	Totals
Boys	190	200	80	30	500
Girls	220	200	60	20	500
Totals	410	400	140	50	1000
Proportion of Girls	.537	.500	.429	.400	.500

Regarding this as a 2 x 4 contingency table with three degrees of freedom, we calculate $\chi^2 = 7.52$, the probability of which value being exceeded by chance is .0570. The indications of a significant difference in distribution of grades between sexes may thus, if one holds to the .05 standard and uses only the χ^2 test, be regarded as not quite significant. There is, however, additional evidence in the fact that the proportion of girls diminishes steadily as we pass down the scale of grades. If we treat excellence in the subject as one variate and the proportion of girls in a group as another, we have a rank correlation of unity, with a sample of four. The probability of a correlation of ± 1 is .083, which also, by itself, would not be considered significant. But we may combine the two pieces of evidence by the method given by Fisher.[8] The process consists of adding the natural logarithms of the two probabilities, doubling, and treating the result as having the χ^2 distribution with four degrees of freedom. This gives a probability in the neighborhood of .03, which would be judged significant.

Similar cases are very common. The value of χ^2 is unchanged if the columns are permuted in any way, whereas r' depends solely on which of the possible permutations actually exists. Thus the two tests are *independent*, a property needed for the combination by the above method.

9. Efficiency of Replacement of Measures by Ranks, and the Estimation of ρ from Rank Correlation, for a Normal Population

Consider a population with a normal distribution in two variates x and y, each of which we shall without loss of generality assume to be of unit variance and zero mean. The density distribution is then specified by $z\,dx\,dy$, where

[8] R. A. Fisher, *Statistical Methods for Research Workers*, 4th and 5th editions, Art. 21.1.

$$z = \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)} \tag{17}$$

where ρ is the correlation of x and y, or the variate correlation. By ξ and η, as in §3, we denote the uniformized variates defined by (5), i.e., functions respectively of x and y having distributions of uniform density from $-\frac{1}{2}$ to $+\frac{1}{2}$. Then ξ and η will each have the variance 1/12. The rank correlation ρ' in the population is the correlation of ξ and η; consequently

$$\rho' = 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\,\eta\, z\, dx\, dy\,. \tag{18}$$

Thus ρ' is a function of ρ, which obviously vanishes when $\rho = 0$.

From (17) the identity

$$\frac{\partial z}{\partial \rho} = \frac{\partial^2 z}{\partial x\, \partial y} \tag{19}$$

is readily calculated. With its help we have from (18) and integrations by parts,

$$\begin{aligned}\frac{d\rho'}{d\rho} &= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\eta\,\frac{\partial z}{\partial \rho}\, dx\, dy = 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\eta\,\frac{\partial^2 z}{\partial x\,\partial y}\, dx\, dy \\ &= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{d\xi}{dx}\frac{d\eta}{dy}\, z\, dx\, dy\,.\end{aligned} \tag{20}$$

Now since x and y are normally distributed with unit variance and zero means, the uniformized variates (5) take the form

$$\xi = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}\, dt\,, \qquad \eta = \frac{1}{\sqrt{2\pi}}\int_0^y e^{-\frac{t^2}{2}}\, dt\,.$$

Therefore

$$\frac{d\xi}{dx} = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}, \qquad \frac{d\eta}{dy} = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}.$$

Substituting these values and (17) in the last integral in (20) we have,

$$\frac{d\rho'}{d\rho} = \frac{12}{4\pi^2\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{(2-\rho^2)x^2-2\rho xy+(2-\rho^2)y^2}{2(1-\rho^2)}}\, dx\, dy\,.$$

The double integral, as is well known, equals π divided by the square root of the discriminant of the quadratic form in the exponent. This gives

$$\frac{d\rho'}{d\rho} = \frac{6}{\qquad}$$

Therefore, since ρ' vanishes with ρ,

$$\rho' = \frac{6}{\pi} \sin^{-1} \frac{\rho}{2},$$

or

$$\rho = 2 \sin \frac{\pi \rho'}{6}.$$

This is essentially the process used by Pearson.

The last equation suggests that an estimate r'' of ρ be based on the rank correlation r' by means of the relation

$$r'' = 2 \sin \frac{\pi r'}{6}.$$

Prefixing a δ to denote a deviation of sample from population value we have by a Taylor expansion,

$$\delta r'' = \frac{\pi}{3} \cos \frac{\pi \rho'}{6} \delta r' + \cdots,$$

the terms dropped being of higher order in $\delta r'$ than those written, and consequently of higher order in n^{-1}. Squaring, taking the expectation, and ignoring the terms of higher order, we have for the case $\rho = \rho' = 0$, by (10),

$$\sigma_{r''}^2 = E(\delta r'')^2 = \frac{\pi^2}{9} \sigma_{r'}^2 = \frac{\pi^2}{9(n-1)},$$

approximately.

The last result enables us to measure the loss of information, at least for large samples, that results from neglecting the exact values of the variates and using only ranks. The product-moment correlation coefficient r has, if $\rho = 0$, the exact variance

$$\frac{1}{n-1},$$

the ratio of which to $\sigma_{r'}^2$ tends as n increases to $9/\pi^2$. Thus the efficiency of the rank correlation method in estimating ρ, if ρ is really zero, is $9/\pi^2 = .9119$. This means that the product-moment correlation is approximately as sensitive a test of the existence of a relationship in a normally distributed population with 91 cases as the rank correlation with 100 cases.

The efficiency of r' will of course be different for non-normal populations, and also for normal populations with $\rho \neq 0$. But if the form of the population is known, this knowledge may always be used to supplement the ranks to obtain a more accurate estimate of correlation, or test of relationship. This fact deserves some attention, since a superficial observation of the coincidence of the formula (1)

for the leading term of the variance of an arbitrary uncorrelated population, and the leading term of the formula (10) for the variance of the rank correlation, might suggest that r' is as accurate as r. But it may be surmised that the 9 % loss of information found for the bivariate normal distribution is the greatest loss of information in using r' in place of r to test for independence, since for non-normal populations the most efficient estimate of the correlation will not usually be r, but a more complicated function of the observations. Certainly where there is complete absence of knowledge of the form of the bivariate distribution, and especially if it is believed not to be normal, the rank correlation coefficient is to be strongly recommended as a means of testing the existence of relationship.

COLUMBIA UNIVERSITY.

THE ELIMINATION OF PERPETUAL CALENDARS

By John L. Roberts

If we wish to find the day of the week for any date, one way to solve the problem is to use a perpetual calendar. Another way to solve the problem is to calculate the day of the week by mathematical methods. In the past these mathematical methods have been so complicated that it has been much more convenient to use a perpetual calendar. This explains why some people have put themselves to the expense of buying perpetual calendars. The purpose of this article is to provide a mathematical method which is so simple that the entire calculation can be done mentally and which is as convenient as a perpetual calendar. In this article this mathematical method is applied to the Gregorian, Julian, and World calendars. Since a great many records have been made using the Julian and Gregorian calendars, the adoption of the World calendar would not completely eliminate the usefulness of applying the mathematical method to the historical calendars. The mathematical method also shows to what extent the World calendar is a simplification; this is important because proposals to reform the present calendar are attracting world-wide attention.

In the theory of numbers occurs the expression,

$$a \equiv b \bmod p, \tag{1}$$

which is read a is congruent to b modulo p, and which means that the difference of a and b is divisible by p. Since p in this article is always equal to 7, it is convenient to represent (1) by

$$a \equiv b. \tag{2}$$

Assume m stands for any number which represents any monthday of any month. Assume w stands for any number which represents any day of the week. It is assumed that 7 stands for Sunday, 1 for Monday, 2 for Tuesday, etc. It is assumed that the constant c for any month is the value of m at the first Sunday in that month. Then (2) becomes

$$w \equiv m - c, \tag{3}$$

which enables us to find w if m is known provided the constant c is known for the month in question. Consequently, all we need to complete our theory is to discover a method of finding c for any possible month.

First, there will be discussed rules for finding c for any month of the Gregorian calendar in 1935. An inspection of the calendar shows that c for December is

equal to 1. Since November has 30 days, we can find c for it by adding 2, which is congruent to 30, to the c for December. Since the number of days in September, October, and November is 91, which is congruent to zero, the c's for September and December have the same value. In like manner, since c for September is 1, the c for June is 2, and the c for March is 3. We now have all the theory which is necessary to find w at any date in 1935. For example, suppose we wish to find w for April 17, and know that the c for December is 1. Then, by adding 2 we find that the c for March is 3. We are now in position easily to calculate that the c for April is 7. Applying (3) we find that w at April 17 is 3, which stands for Wednesday.

All that is necessary to complete our theory of the Gregorian calendar is to find rules for finding c for December of any possible year, because, if this is known, we can find c for any month in that year by the method used for 1935. It is convenient to represent the expression, "c for December 1935" by "C for 1935." In like manner C for any calendar year means c for December of that year. Since C for 1935 is 1 and since the number of days in 1936 is 366, which is congruent to 2, subtracting this 2, we find that C for 1936 is 6, because -1 is congruent to 6. Knowing C for 1936, we deduce that C for 1940, which is four years later, is 1, because $6 + 2$ is congruent to 1; and that C for 1928 is 2, found by subtracting 4. The C's for 1900, 1928, 1956, and 1984 are equal. Full centuries in order to be leap years must be divisible by 400. Since C for 1900 is 2, we find by adding 1 that C for 2000 is 3. Knowing C for 2000, we deduce by adding 2 that C for 2100 is 5. 1600, 2000, and 2400 have the same value of C. If it is assumed that the length of the tropical year is exactly 365.2425 days, we have all the theory which is necessary to find C for any possible year. Although this assumption contains a small error, any further discussion of it would hardly be of any practical interest. The foregoing theory provides complete methods for finding w by means of a series of steps, which are so simple that the entire calculation can be done mentally. For example, suppose we wish to find w for November 29, 1888. Each of the C's for 1800 and 1884 is 7. Therefore, C for 1888 is 2, which is congruent to $7 + 2$. Adding 2, c for November of this year is 4. Applying (3), we find that w at November 29, 1888 is 4, which stands for Thursday. In order to calculate mentally w for any date of the Gregorian calendar, it is only necessary for me to remember the foregoing mathematical method and to remember I was born on November 29, 1888, a Thanksgiving Day.

Deplorable changes were made in the Julian calendar between 45 B.C. and 1 A.D. Also it was not until 325 A.D. that the use of the 7-day week became general throughout the Roman Empire, gradually supplanting the old division of the month into Calends, Nones, and Ides. Therefore, in order to save space, the application of our theory prior to 1 A.D. is left to the reader. Starting with this year it is only necessary to discover a rule for finding the C's of the Julian calendar for the full centuries, because the rules of the Gregorian calendar apply to all other years. October 5, 1582, Old Style was the same day as Oc-

tober 15, 1582, New Style; the Gregorian calendar was born at this date. December 17, 1600, New Style was a Sunday, and was the same day as December 7, 1600, Old Style. Therefore, C for 1600, Old Style is 7. It is now a very simple matter to complete our theory of the Julian calendar. Since C for 1600 is 7, subtracting 1, C for 1500 is 6. 200, 900, and 1600 have the same value of C.

In the case of the World calendar the c's for the three months of each of the equal quarters can be found as follows. For the first month c is 1. Therefore, c for the second month is 5, which is congruent to $1 - 3$. Subtracting 2 from this 5, we find that c for the third month is 3.

NOTES

ON STANDARD ERROR FOR THE LINE OF MUTUAL REGRESSION

By Y. K. Wong

1. In Pearson's *On Lines and Planes of Closest Fit to System of Points in Space*, he establishes a formula for the mean square residual for the best fitting line in q-space:

$$(1) \qquad (\text{mean sq. residual})^2 = \sigma_{x_1}^2 + \cdots + \sigma_{x_q}^2 - \Delta R_{\max}^2$$

where $2R_{\max}$ is the length of the maximum axis of the correlation ellipse in q-space, and Δ is the correlation determinant.[1]

In the present paper, we consider a 2-dimensional case, and shall call the mean sq. residual as the standard error, denoted by S_N.

In 2-dimensional space, a correlation ellipse is

$$(2) \qquad ax^2 + 2hxy + by^2 + c = 0,$$

where

$$(2a) \qquad a = \sigma_y^2, \quad b = \sigma_x^2, \quad h = -r_{xy}\sigma_x\sigma_y = -p_{xy} = -p_{yx}, \quad c = -\sigma_x^2\sigma_y^2.$$

Pearson gives in the 2-dimensional space the following formula for S_N:

$$(3) \qquad S_N = \sigma_x\sigma_y/\text{semi-major axis of equation (2)}.$$

Expression (3) can be readily deduced from (1). This paper aims to present some formulae for S_N, more convenient for practical computation, and also call attention to a misprint in Pearson's paper.

2. From analytic geometry, we see that the angle φ, between the major axis of the ellipse (2) and the x-axis is given by

$$(4) \qquad \tan 2\varphi = 2h/(a - b).$$

By rotation of the axes, equation (1) can be written in the form

$$(5) \qquad a'x^2 + b'y^2 + c = 0,$$

where

$$(5a) \qquad \begin{aligned} a' &= a\cdot\cos^2\varphi - 2h\cdot\sin\varphi\cdot\cos\varphi - b\cdot\sin^2\varphi > 0 \\ b' &= a\cdot\sin^2\varphi - 2h\cdot\sin\varphi\cdot\cos\varphi - b\cdot\cos^2\varphi > 0. \end{aligned}$$

[1] Philosophical Magazine, 6th Series, II (November, 1901), p. 559.

LEMMA 1. The value of a' given by (5a) is less than b'.

To prove this lemma, we find from (4) and (5)

$$a' - b' = a + b, \qquad a' - b' = 2h/\sin 2\varphi = -2p_{xy}/\sin 2\varphi,$$

and hence

$$2a' = a + b - 2p_{xy}/\sin 2\varphi, \qquad 2b' = a + b + 2p_{xy}/\sin 2\varphi. \tag{6}$$

Since both a and b are positive, the lemma will be proved if we can show that $p_{xy}/\sin 2\varphi$ is a positive quantity. By (2a), $p_{xy} = r_{xy}\sigma_x\sigma_y$, in which σ_x, σ_y are positive; hence the sign of p depends upon the sign of r. If $r_{xy} < 0$, then $\varphi > \frac{\pi}{2}$, and 2φ is of such a nature that $\frac{3\pi}{2} < 2\varphi < 2\pi$. It follows $\sin 2\varphi < 0$, and hence $p_{xy}/\sin 2\varphi$ is positive. On the other hand, if $r_{xy} > 0$, then φ is such that $0 < 2\varphi < \pi$, and hence $\sin 2\varphi > 0$. It follows that $p_{xy}/\sin 2\varphi$ is positive independent of the sign of r_{xy}.

LEMMA 2. The square of the mean square residual is equal to a', and hence

$$S_N^2 = \sigma_y^2 \cos^2\varphi - 2p_{xy}\sin\varphi\cos\varphi + \sigma_x^2\sin^2\varphi = \tfrac{1}{2}(\sigma_x^2 - \sigma_y^2) - p_{xy}/\sin 2\varphi.$$

For from (5), we obtain $(\text{semi-major axis})^2 = -c/a' = +\frac{\sigma_x^2\sigma_y^2}{a'}$. Substituting this into (3), we obtain $S_N = a'$. The balance of the lemma follows from (5a), (6), and (2a).

LEMMA 3. For every r_{xy}, we have

$$\sin 2\varphi = p_{xy}/\sqrt{K}, \qquad K = (\sigma_x^2 - \sigma_y^2)^2 + 4p_{xy}^2. \tag{7}$$

For, from (4), we find $\sin 2\varphi = -p_{xy}/\pm\sqrt{K} = r_{xy}\left(\frac{-\sigma_x\sigma_y}{\pm\sqrt{K}}\right)$. By the argument given in the demonstration of Lemma 1, we see that r_{xy} and $\sin 2\varphi$ should be of the same sign. Hence the negative sign is chosen before the radical.

From Lemma 2 and (7), we have the formula given by Pearson:

$$2S_N^2 = (\sigma_x^2 + \sigma_y^2)^2 - \sqrt{K}. \tag{8}$$

3. We are going to establish several more formulae for S_N. From (4), we have $2h\cdot\tan(\varphi) = -(a - b) \pm \sqrt{K}$. The sign before the radical is determined in such a way that $\tan(\varphi)$ has the same sign as r_{xy}. By the reasoning given in Lemma 1, the negative sign is chosen. Thus

$$-2p_{xy}\cdot\tan\varphi \mp -(\sigma_y^2 - \sigma_x^2) - \sqrt{K} = \sigma_x^2 + \sigma_y^2 - \sqrt{K} - 2\sigma_y^2$$

or

$$2(\sigma_y^2 - p_{xy}\tan\varphi) = \sigma_x^2 - \sigma_y^2 - \sqrt{K}.$$

This proves that

$$S_N^2 = \sigma_y^2 - p_{xy}\tan\varphi. \tag{9}$$

Similarly, we have

$$S_N^2 = \sigma_x^2 - p_{xy} \cot \varphi. \tag{10}$$

For computation, (9) and (10) are more convenient than (8). When the line of mutual regression is determined, it is known that $\tan \varphi$ (denoted by B) is equal to the slope of that line, and hence $\cot \varphi$ $(= 1/B)$ is equal to the reciprocal of the slope. Then we can write (9) and (10) as follows:

$$S_N^2 = \sigma_y^2 - p_{yx} \cdot B \tag{11}$$

$$S_N^2 = \sigma_x^2 - p_{xy}/B. \tag{12}$$

The second formula given in Lemma 2 is simpler than (8), but not as simple as (11) and (12).

For computation, it is convenient to find φ from the equation

$$\tan 2\varphi = \frac{+2r_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} = H,$$

i.e.,

$$2\varphi = \arctan H.$$

Since $\sin 2\varphi$ and r_{xy} are of the same sign, we can determine the value of φ from the preceding equation by inspection, though $\arctan H$ is a multiple-valued function. After the determination of φ, we can obtain

$$B = \tan \varphi.$$

Then we can compute S_N either from (9), (11), or (10), (12).

There is a very interesting fact furnished by (11) and (12). These two formulae are, in fact, generalizations of the following two well known ones:

$$S_y^2 = \sigma_y^2(1 - r) \tag{a}$$

$$S_x^2 = \sigma_x^2(1 - r), \tag{b}$$

where S_y is the standard error of the line of regression when y is used as dependent variable and x as independent variable, and similarly for S_x. It is clear that the line of mutual regression may be looked upon as a generalization of the other two lines of regression when we use y or x as dependent variable. So the slope B of the line of mutual regression is a generalization of $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$. where the subscript yx means y on x and xy, x on y. If we use x as independent variable, then we must obtain b_{yx} instead of B. Hence substituting the formula of b_{yx} instead of B into (11), we obtain, after a simple reduction, the same result as given by (a). On the other hand, if we use y as independent variable, we must obtain b_{xy} instead of $1/B$. It will result (b) when b_{xy} is put in the place of $1/B$ in (12). The generalization perhaps can be seen more clearly if we write (a) and (b) into slightly different forms:

(a′) $$S_y^2 = \sigma_y^2 - p_{yx} \cdot b_{yx}$$

(b′) $$S_x^2 = \sigma_x^2 - p_{xy} \cdot b_{xy}.$$

4. The misprint in Pearson's paper is on the second formula of the following:

$$(MSR)^2 = \frac{\sigma_x^2 \sigma_y^2}{\cot^2 \varphi} = \frac{1}{2}\left(\sigma_x^2 - \sigma_y^2 - \sqrt{(\sigma_x^2 - \sigma_y^2)^2 - 4r^2 \sigma_x^2 \sigma_y^2}\right)$$

where $\tan 2\varphi = 2r_{xy}\sigma_x\sigma_y/(\sigma_x^2 - \sigma_y^2)$. $\cot^2 \varphi$ should read "square of semi-major axis of ellipse (2)." Professor Henry Schultz first noticed this misprint and suggested to the writer to investigate it.

In a recent letter to Schultz, Pearson pointed out that one of the simplest formula for S_N^2 or $(MSR)^2$ is given by

(α) $$S_N^2 = \sigma_x^2 \sin^2 \varphi + \sigma_y^2 \cos^2 \varphi,$$

where φ is defined by (4). However, Professor Schultz expressed doubt about its validity. From lemma 2, it is clear that (α) is also not true.

INSTITUTE OF SOCIAL SCIENCES,
ACADEMIA SINICA, PEIPING

THE DISTRIBUTION LAWS OF THE DIFFERENCE AND QUOTIENT OF VARIABLES INDEPENDENTLY DISTRIBUTED IN PEARSON TYPE III LAWS[1]

BY SOLOMON KULLBACK

Although the results herein described are not entirely new, it is felt that the method of solution is of interest as presenting further illustrations of the application of characteristic functions to the distribution problem of statistics (*1*).

1. Distribution law of the difference. Let $u = x - y$, where the distribution laws of x and y are independent and given respectively by

$$f_1(x) = \frac{e^{-x}x^{p-1}}{\Gamma(p)}; \quad f_2(y) = \frac{e^{-y}y^{q-1}}{\Gamma(q)} \qquad 0 \leqq x \leqq \infty; 0 \leqq y \leqq \infty. \tag{1}$$

The characteristic function of the distribution law of u is given by (*1*),

$$\varphi(t) = \int_0^\infty \frac{e^{itx-x}x^{p-1}\,dx}{\Gamma(p)} \int_0^\infty \frac{e^{-ity-y}y^{q-1}\,dy}{\Gamma(q)} \tag{2}$$

$$= \frac{1}{(1-it)^p(1+it)^q}. \tag{3}$$

The distribution law of u is given by (*1*),

$$D(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itu}\,dt}{(1-it)^p(1+it)^q}. \tag{4}$$

Let $1 - it = -\frac{z}{u}$,

$$D(u) = \frac{e^{-u}u^{p-1}}{2^q\,2\pi i}\int_{-u-i\infty}^{-u+i\infty} \frac{e^{-z}\,dz}{(-z)^p\left(1+\frac{z}{2u}\right)^q}. \tag{5}$$

Now it may be shown that (*1*)

$$\frac{1}{2\pi i}\int_{-u-i\infty}^{-u+i\infty} \frac{e^{-z}\,dz}{(-z)^p\left(1+\frac{z}{2u}\right)^q} = \frac{e^u(2u)^{\frac{q-p}{2}}}{\Gamma(p)} W_{\frac{p-q}{2},\frac{1-p-q}{2}}(2u) \tag{6}$$

[1] Presented to the American Mathematical Society, June 20, 1934.

where $W_{k,m}(z)$ is the confluent hypergeometric function (*2*). Since $W_{k,m}(z) = W_{k,-m}(z)$ we have finally

$$D(u) = \frac{u^{\frac{p+q}{2}-1}}{2^{\frac{p+q}{2}}\Gamma(p)} W_{\frac{p-q}{2},\frac{p+q-1}{2}}(2u). \tag{7}$$

For $p = q$, since $W_{0,m}(2x) = \frac{x^{\frac{1}{2}}2^{\frac{1}{2}}}{\sqrt{\pi}} K_m(x)$ where $K_m(x)$ is the Bessel Function of the second kind and imaginary argument (*1*), we obtain

$$D(u) = \frac{u^{\frac{2p-1}{2}}}{2^{\frac{2p-1}{2}}\Gamma(p)\sqrt{\pi}} K_{\frac{2p-1}{2}}(u). \tag{8}$$

This result has been otherwise obtained by Pearson, Stouffer, and David (*3*).

2. Distribution law of the quotient. Let $u = \log x - \log y$ where x and y are defined as above.

The characteristic function of the distribution law of u is given by (*1*)

$$\varphi(t) = \int_0^\infty \frac{e^{-x}x^{p-1+it}\,dx}{\Gamma(p)} \int_0^\infty \frac{e^{-v}y^{q-1-it}\,dy}{\Gamma(q)} \tag{9}$$

$$= \frac{\Gamma(p+it)\,\Gamma(q-it)}{\Gamma(p)\,\Gamma(q)}. \tag{10}$$

The distribution law of u is given by (*1*)

$$D(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itu}\,\Gamma(p+it)\,\Gamma(q-it)}{\Gamma(p)\,\Gamma(q)}\,dt. \tag{11}$$

Let $q - it = -z$, so that

$$D(u) = \frac{e^{-qu}}{\Gamma(p)\,\Gamma(q)\,2\pi i}\int_{-q-i\infty}^{-q+i\infty} e^{-zu}\,\Gamma(p+q+z)\,\Gamma(-z)\,dz. \tag{12}$$

Now it may be shown that (*2*)

$$\frac{1}{2\pi i}\int_{-q-i\infty}^{-q+i\infty} e^{-zu}\,\Gamma(p+q+z)\,\Gamma(-z)\,dz = \Gamma(p+q)\,(1+e^{-u})^{-(p+q)},$$

so that

$$D(u) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)}\,\frac{e^{pu}}{(1+e^u)^{p+q}}. \tag{13}$$

Since $e^u = \frac{x}{y} = w$, we obtain as the distribution law of the quotient

$$p(w) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)} \frac{w^{p-1}}{(1+w)^{p+q}}. \tag{14}$$

If in (13) we set

$$p = \frac{n_1}{2}; \qquad q = \frac{n_2}{2}; \qquad e^u = \frac{n_1}{n_2} e^{2z},$$

we obtain

$$D(z) = \frac{2\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \frac{n_1^{\frac{n_1}{2}} n_2^{\frac{n_1}{2}} e^{n_1 z}}{(n_2 + n_1 e^{2z})^{\frac{n_1+n_2}{2}}} \tag{15}$$

which result has been otherwise obtained by R. A. Fisher (*4*).

GEORGE WASHINGTON UNIVERSITY.

REFERENCES

(1) KULLBACK, S.: An application of characteristic functions to the distribution problem of statistics. Annals of Mathematical Statistics, Vol. 5 (1934) pp. 263–307.

(2) WHITTAKER AND WATSON: Modern Analysis, 2nd Ed., pp. 333, 283.

(3) PEARSON, STOUFFER AND DAVID: Further applications in statistics of the Bessel Function. Biometrika, Vol. 24 (1932), pp. 293.

(4) FISHER, R. A.: On a distribution yielding the error functions of several well known statistics. Proceedings International Mathematical Congress, Toronto (1924), Vol. 2, pp. 805–813.

REPORT OF THE MEETING OF THE INSTITUTE OF MATHEMATICAL STATISTICS AT ST. LOUIS

The Institute of Mathematical Statistics held a joint meeting with the Econometric Society and the American Mathematical Society at St. Louis, Missouri, on January 2, 1936. The program consisted of an invited address, "The Mathematical Theory of Index Numbers," by Professor Thomas Rawles, and the following additional papers:

(1) On Certain Distributions Derived from the Multinomial Distribution, by Dr. Solomon Kullback

(2) Convexity Properties of Generalized Mean Value Functions, by Dr. Nilan Norris

(3) The Frequency Distribution for the Mean of n Independent Chance Variables When Each Is Subject to the Law $y_0 x^{p-1}(1-x)^{q-1}$, by Prof. W. D. Baten

(4) On the Admissibility of Time Series, by Prof. Francis Regan

The Institute voted to hold a meeting at Cambridge, Massachusetts, early in September of this year. This meeting will be in connection with the celebration of the Harvard Tercentenary. Professor R. A. Fisher will deliver an invited address before the Institute and the American Mathematical Society. A more detailed announcement of the meeting will be made later.

SHEPPARD'S CORRECTIONS FOR A DISCRETE VARIABLE

CECIL C. CRAIG

In the Annals of Mathematical Statistics,[1] J. R. Abernethy gave a derivation of the corrections to eliminate the systematic errors in the moments of a discrete variable due to grouping. It is the purpose of this note to considerably shorten and simplify the derivation of these corrections by an adoption of a device used by R. A. Fisher (not published so far as I know) in the case of the ordinary Sheppard's corrections.

Let us suppose that m consecutive values of the discrete variable in question are grouped in a frequency class of width k. The m smaller intervals of width k/m go to make up the class width k, the actual points representing the m values of the variable being plotted at the centers of the sub-intervals. Now let us suppose that each of m consecutive boundary points of the sub-intervals is as likely to be chosen as a boundary point of the larger intervals as any other. Then, if x_i is the class mark of the i-th frequency class, for any *true* value, x, of the discrete variable included in this frequency class, we have

$$x_i = x + \epsilon$$

in which x and ϵ are independent variables and ϵ takes on the m values

$$-\frac{m-1}{2}k/m,\ -\frac{m-3}{2}k/m,\ \cdots,\ \frac{m-3}{2}k/m,\ \frac{m-1}{2}k/m,$$

with the equal relative frequencies $1/m$.

The moments of x_i are those calculated from the grouped frequency distribution; the problem is to express the average values of the moments of x in terms of the calculated moments and k and m. The use of moment generating functions at once leads to the desired results. Denoting the s-th moment of x_i about any origin by ν_s', the like moment of x by μ_s', the respective moment generating functions of the two variables by $M_{x_i}(\vartheta)$ and $M_x(\vartheta)$ respectively, we have at once

$$M_{x_i}(\vartheta) = M_x(\vartheta) \sum_{\epsilon=-\frac{m-1}{2}k/m}^{\frac{m-1}{2}k/m} \frac{e^{\epsilon\vartheta}}{m}, \tag{1}$$

[1] "On the Elimination of Systematic Errors Due to Grouping," vol. IV (1933), pp. 263–277.

in which by definition

$$M_{x_i}(\vartheta) = 1 + \nu_1'\vartheta + \nu_2'\,\vartheta^2/2! + \nu_3'\,\vartheta^3/3! + \cdots,$$
$$M_x(\vartheta) = 1 + \mu_1'\vartheta + \mu_2'\,\vartheta^2/2! + \mu_3'\,\vartheta^3/3! + \cdots.$$

The computation necessary to get the actual corrections consists in the calculation of the coefficients in the formal expansion of

$$M_\epsilon(\vartheta) = \sum_{\epsilon=-\frac{m-1}{2}\,k/m}^{\frac{m-1}{2}\,k/m} \frac{e^{\epsilon\vartheta}}{m}, \tag{2}$$

in powers of ϑ and then solving for the μ_s''s in (1).

But the summation indicated in (2) is readily effected by means of the calculus of finite differences. In fact, we get

$$M_\epsilon(\vartheta) = \frac{e^{\frac{m+1}{2}\,k\vartheta/m} - e^{-\frac{m-1}{2}\,k\vartheta/m}}{m(e^{k\vartheta/m} - 1)} = \frac{\sinh k\vartheta/2}{m \sinh k\vartheta/2m}. \tag{3}$$

Then (2) becomes

$$M_{x_i}(\vartheta) = M_x(\vartheta)\,\frac{\sinh k\vartheta/2}{m \sinh k\vartheta/2m}. \tag{4}$$

If we let $m \to \infty$ we get the corresponding result for a continuous variable

$$M_{x_i}(\vartheta) = M_x(\vartheta)\,\frac{\sinh k\vartheta/2}{k\vartheta/2} \tag{5}$$

already given by Langdon and Ore,[2] though in a less elegant manner; for in this case, the expression analogous to (1) is immediately seen to be

$$M_{x_i}(\vartheta) = M_x(\vartheta) \int_{-k/2}^{k/2} e^{\epsilon\vartheta}\,\vartheta\epsilon/k\,.$$

Returning to (4), taking the logarithms of both sides, remembering that the logarithm of the moment generating function is the generating function of the semi-invariants of Thiele, we get,

$$\lambda_1\vartheta + \lambda_2\,\vartheta^2/2! + \lambda_3\,\vartheta^3/3! + \cdots$$
$$= \bar{\lambda}_1\vartheta + \bar{\lambda}_2\,\vartheta^2/2! + \bar{\lambda}_3\,\vartheta^3/3! + \cdots - \log \frac{k\vartheta/2m \sinh k\vartheta/2}{k\vartheta/2 \sinh k\vartheta/2m}, \tag{6}$$

in which the $\bar{\lambda}_r$'s are the calculated semi-invariants and the λ_r's the corrected ones.

[2] W. H. Langdon and O. Ore, Semi-invariants and Sheppard's Corrections, Annals of Mathematics, vol. 31 (1930), pp. 230–232.

But since

$$\log \frac{\sinh x}{x} = \sum_{s=1}^{\infty} (-1)^{s+1} \frac{B_s}{2s\,(2s)!} (2x)^{2s}$$

we have on setting:

$$(7) \qquad -\log \frac{k\vartheta/2m \sinh k\vartheta/2}{k\vartheta/2 \sinh k\vartheta/2m} = a_0 + a_1\vartheta + a_2\,\vartheta^2/2! + a_3\,\vartheta^3/3! + \cdots,$$

$$(8) \qquad \begin{aligned} a_0 &= 0, \qquad a_{2s+1} = 0, \qquad s = 0, 1, 2, \cdots \\ a_{2s} &= \frac{(-1)^s B_s k^{2s}}{2s}\left(1 - \frac{1}{m^{2s}}\right), \qquad s = 1, 2, 3, \cdots. \end{aligned}$$

Obviously these a's are the "Sheppard's" corrections for the semi-invariants. We have generally

$$\begin{aligned} \lambda_{2s+1} &= \bar{\lambda}_{2s+1}, \qquad s = 0, 1, 2, \cdots. \\ \lambda_{2s} &= \bar{\lambda}_{2s} + (-1)^s \frac{B_s k^{2s}}{2s}\left(1 - \frac{1}{m^{2s}}\right). \end{aligned}$$

In particular

$$\lambda_2 = \bar{\lambda}_2 - \left(1 - \frac{1}{m^2}\right) k^2/12 \qquad \lambda_6 = \bar{\lambda}_6 - \left(1 - \frac{1}{m^6}\right) k^6/252$$

$$\lambda_4 = \bar{\lambda}_4 + \left(1 - \frac{1}{m^4}\right) k^4/120 \qquad \lambda_8 = \bar{\lambda}_8 + \left(1 - \frac{1}{m^8}\right) k^8/240.$$

For $m \to \infty$, these give of course the results reached by Langdon and Ore.[3]

To get the corrections for the moments let us set

$$\frac{m \sinh k\vartheta/2m}{\sinh k\vartheta/2} = \alpha_0 + \alpha_1\vartheta + \alpha_2\,\vartheta^2/2! + \alpha_3\,\vartheta^3/3! + \cdots.$$

From (7) and (8)

$$(9) \qquad \begin{aligned} \alpha_0 &= 1, \qquad \alpha_{2n+1} = 0, \qquad n = 0, 1, 2, \cdots \\ \alpha_{2n} &= \sum \frac{(2n)!\, a_2^r a_4^s a_6^t \cdots}{(2!)^r (4!)^s (6!)^t \cdots r!\, s!\, t! \cdots} \end{aligned}$$

the summation extending over all positive, integral values of $r, s, t, \cdots$ for which,

$$r + 2s + 3t + \cdots = n.$$

---

[3] Loc. cit.

Then finally we have the formula,

$$\mu_n' = \sum_{s=0}^{\left[\frac{n}{2}\right]} \binom{n}{2s} \alpha_{2s} \nu'_{n-2s}, \tag{10}$$

for the corrected moments.

Writing out the first four α's, we have for the first eight moments about the mean

$$\mu_1 = \nu_1 = 0$$

$$\mu_2 = \nu_2 - (1 - 1/m^2)\, k^2/12$$

$$\mu_3 = \nu_3$$

$$\mu_4 = \nu_4 - (1 - 1/m^2)\, \nu_2 k^2/2 + (1 - 1/m^2)(7 - 3/m^2)\, k^4/240.$$

$$\mu_5 = \nu_5 - 5(1 - 1/m^2)\nu_3 k^2/6.$$

$$\mu_6 = \nu_6 - 5(1 - 1/m^2)\, \nu_4 k^2/4 + (1 - 1/m^2)(7 - 3/m^2)\, \nu_2 k^4/16 - (1 - 1/m^2)(31 - 18/m^2 + 3/m^4)\, k^6/1344$$

$$\mu_7 = \nu_7 - 7(1 - 1/m^2)\, \nu_5 k^2/4 + 7(1 - 1/m^2)(7 - 3/m^2)\, \nu_3 k^4/48$$

$$\mu_8 = \nu_8 - 7(1 - 1/m^2)\, \nu_6 k^2/3 + 7(1 - 1/m^2)(7 - 3/m^2)\, \nu_4 k^4/24 - (1 - 1/m^2)(31 - 18/m^2 + 3/m^4))\, \nu_2 k^6/48 + (1 - 1/m^2)(381 - 239/m^2 + 55/m^4 - 5/m^6)\, k^8/11520.$$

The final term in μ_{2n} as given above is α_{2n}.

The above method is readily extended to the case of two or more variables. We will illustrate the procedure by getting the results likely to be required for two variables. As before we suppose that m consecutive values of x are grouped in a frequency class of width k, and we shall similarly suppose that n values of y are grouped in a frequency class of width l. And arguing as before we write now

$$x_i = x + \epsilon$$

$$y_j = y + \eta$$

in which ϵ and η are independent of x and y and of each other.

The moment generating function of two variables is defined by the identity in ϑ and ω:

$$M_{x,y}(\vartheta, \omega) = 1 + (\mu'_{10}\vartheta + \mu'_{01}\omega) + \frac{1}{2!}(\mu'_{20}\vartheta^2 + 2\mu'_{11}\vartheta\omega + \mu'_{02}\omega^2) + \cdots$$

$$= 1 + (\mu'_{10}\vartheta + \mu'_{01}\omega) + \frac{1}{2!}(\mu'_{10}\vartheta + \mu'_{01}\omega)^{(2)} + \frac{1}{3!}(\mu'_{10}\vartheta + \mu'_{01}\omega)^{(3)} + \cdots,$$

in which the manner of expansion of $(\mu'_{10}\vartheta + \mu'_{01}\omega)^{(r)}$ is evident.

Then from the properties of moment generating functions, we have

$$M_{x_i,\,x_j}(\vartheta,\omega) = M_{x,\,y}(\vartheta,\omega) \sum_{\epsilon=-\frac{m-1}{2}\ k/m}^{\frac{m-1}{2}\ k/m} \sum_{\eta=-\frac{n-1}{2}\ l/n}^{\frac{n-1}{2}\ l/n} \frac{e^{\epsilon\vartheta+\eta\omega}}{mn} \tag{11}$$

$$= M_{x,\,y}(\vartheta,\omega)\,\frac{\sinh k\vartheta/2}{m\sinh k\vartheta/2m}\,\frac{\sinh l\omega/2}{n\sinh l\omega/2n}.$$

As in the case of a single variable it will be simpler first to get the corrections for the semi-invariants. The logarithm of the moment generating function is the generating function of the semi-invariants; thus

$$\log M_{x,y}(\vartheta,\omega) = (\lambda_{10}\vartheta + \lambda_{10}\omega) + \frac{1}{2!}(\lambda_{10}\vartheta + \lambda_{01}\omega)^{(2)} + \frac{1}{3!}(\lambda_{10}\vartheta + \lambda_{01}\omega)^{(3)} + \cdots,$$

in which

$$(\lambda_{10}\vartheta + \lambda_{01}\omega)^{(3)} = \lambda_{30}\vartheta^3 + 3\lambda_{21}\vartheta^2\omega + 3\lambda_{12}\vartheta\omega^2 + \lambda_{03}\omega^3,$$

etc.

We write (see (7)),

$$\log\frac{m\sinh k\vartheta/2m}{\sinh k\vartheta/2} = a_2\vartheta^2/2! + a_4\vartheta^4/4! + \cdots$$

$$\log\frac{n\sinh l\omega/2n}{\sinh l\omega/2} = b_2\omega^2/2! + b_4\omega^4/4! + \cdots, \tag{12}$$

with

$$a_{2r} = \frac{(-1)^r B_r k^{2r}}{2r}(1 - 1/m^{2r})$$

$$b_{2s} = \frac{(-1)^s B_s l^{2s}}{2s}(1 - 1/n^{2s}).$$

Then from (11) we have

$$(\lambda_{10}\vartheta + \lambda_{01}\omega)^{(2s+1)} = (\bar\lambda_{10}\vartheta + \bar\lambda_{01}\omega)^{(2s+1)}, \qquad s = 0, 1, 2, \cdots$$

$$(\lambda_{10}\vartheta + \lambda_{01}\omega)^{(2s)} = (\bar\lambda_{10}\vartheta + \bar\lambda_{01}\omega)^{(2s)} + a_{2s}\vartheta^{2s} + b_{2s}\omega^{2s}, \qquad s = 1, 2, 3, \cdots, \tag{13}$$

in which, of course, $\bar\lambda_{rs}$ is a calculated semi-invariant and λ_{rs} a corrected one. We read off

$$\lambda_{rs} = \bar\lambda_{rs}, \qquad rs \neq 0,$$

as already shown by Wold in the case of continuous variables,[4]

$$\lambda_{2s+1,\,0} = \bar\lambda_{2s+1,\,0}, \qquad \lambda_{0,\,2s+1} = \bar\lambda_{0,\,2s+1}.$$

[4] Herman Wold: Sheppard's Correction Formulae in Several Variables: Skandinavisk Aktuarietidskrift, vol. XVII (1934), pp. 248–255.

The values of $\lambda_{2s,0}$ are the same as those for λ_{2s} given above and those for $\lambda_{0,2s}$ are obtained from these merely by replacing in them m and k by n and l. And it is quite obvious that for any number of variables the only semi-invariants to be corrected are those in which a single figure of the index is different from zero and is moreover even. For such semi-invariants the corrections are naturally those derived for a single variable.

Now to derive the corrections for the moments, we write

$$\frac{m \sinh k\vartheta/2m}{\sinh k\vartheta/2} \cdot \frac{n \sinh l\omega/2n}{\sinh l\omega/2} = e^{1/2!\,(a_2\vartheta^2+b_2\omega^2)+1/4!\,(a_4\vartheta^4+b_4\omega^4)+\cdots}$$

$$= 1 + 1/2!\,(\alpha_{20}\vartheta^2 + \alpha_{02}\omega^2) + 1/4!\,(\alpha_{20}\vartheta^2 + \alpha_{02}\omega^2)^{(2)} + \cdots,$$

with now,

$$(\alpha_{20} + \alpha_{02})^{(h)} = \sum \frac{(2h)!\,(a_2 + b_2)^r\,(a_4 + b_4)^s \cdots}{(2!)^r\,(4!)^s \cdots r!\,s! \cdots},$$

the summation to be over all positive integral values of $r, s, \cdots$ for which

$$r + 2s + \cdots = h$$

and in which the parameters ϑ and ω may be omitted without ambiguity.

The formula for the corrected moments can now be written

$$(14) \qquad (\mu'_{10} + \mu'_{01})^{(p)} = \sum_{q=0}^{[p/2]} \binom{p}{2q} (\alpha_{20} + \alpha_{02})^{(q)}\,(\nu'_{10} + \nu'_{01})^{(p-2q)}.$$

This gives

$$\begin{aligned}
&\mu'_{10} + \mu'_{01} = \nu'_{10} + \nu'_{01} \\
&(\mu'_{10} + \mu'_{01})^{(2)} = (\nu'_{10} + \nu'_{01})^{(2)} + (\alpha_{20} + \alpha_{02}) \\
(15)\quad &(\mu'_{10} + \mu'_{01})^{(3)} = (\nu'_{10} + \nu'_{01})^{(3)} + 3(\nu'_{10} + \nu'_{01})\,(\alpha_{20} + \alpha_{02}) \\
&(\mu'_{10} + \mu'_{01})^{(4)} = (\nu'_{10} + \nu'_{01})^{(4)} + 6(\nu'_{10} + \nu'_{01})^{(2)}\,(\alpha_{20} + \alpha_{02}) + (\alpha_{20} + \alpha_{02})^{(2)} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned}$$

Noting that,

$$(\alpha_{20} + \alpha_{02})^{(2)} = a_4 + b_4 + 3(a_2 + b_2)^2,$$

we get the following formulas for the correction of the product moments about an arbitrary origin:

$$\begin{aligned}
&\mu'_{11} = \nu'_{11} \\
&\mu'_{21} = \nu'_{21} - (1 - 1/m^2)\,\nu'_{01}\,k^2/12 \\
&\mu'_{12} = \nu'_{12} - (1 - 1/n^2)\,\nu'_{10}\,l^2/12 \\
&\mu'_{31} = \nu'_{31} - (1 - 1/m^2)\,\nu'_{11}\,k^2/4 \\
&\mu'_{22} = \nu'_{22} - (1 - 1/m^2)\,\nu'_{20}\,l^2/12 - (1 - 1/n^2)\,\nu'_{02}\,k^2/12 \\
&\qquad\qquad - (1 - 1/m^2)\,(1 - 1/n^2)\,k^2l^2/144 \\
&\mu'_{13} = \nu'_{13} - (1 - 1/n^2)\,\nu'_{11}\,l^2/4.
\end{aligned}$$

The above results give the corrections for moments about the mean, merely by dropping the primes and setting $\nu_{10} = \nu_{01} = 0$. In practice the corrections needed are for moments about the mean, and though there would be no difficulty in computing additional results for an arbitrary origin, I shall give here only the additional results for moments about the mean through the sixth order, omitting those obtained merely by permutation of subscripts and interchange of k and m with l and n respectively.

First, the necessary extension of (15) is

$$(\mu_{10} + \mu_{01})^{(5)} = (\nu_{10} + \nu_{01})^{(5)} + 10\,(\nu_{10} + \nu_{01})^{(3)}\,(\alpha_{20} + \alpha_{02})$$

$$(15) \qquad (\mu_{10} + \mu_{01})^{(6)} = (\nu_{10} + \nu_{01})^{(6)} + 15\,(\nu_{10} + \nu_{01})^{(4)}\,(\alpha_{20} + \alpha_{02})$$

$$+ 15\,(\nu_{10} + \nu_{01})^{(2)}\,(\alpha_{20} + \alpha_{02})^{(2)} + (\alpha_{20} + \alpha_{02})^{(3)}\,.$$

We need the additional relation:

$$(\alpha_{20} + \alpha_{02})^{(3)} = a_6 + b_6 + 15(a_4 + b_4)\,(a_2 + b_2) + 15\,(a_2 + b_2)^3\,.$$

The additional formulas for product moments about the mean follow:

$$\mu_{41} = \nu_{41} - (1 - 1/m^2)\,\nu_{21}\,k^2/12$$

$$\mu_{32} = \nu_{32} - (1 - 1/n^2)\,\nu_{30}\,l^2/2 - (1 - 1/m^2)\,\nu_{12}\,k^2/4$$

$$\mu_{51} = \nu_{51} - (1 - 1/m^2)\,5\nu_{31}\,k^2/6 + (1 - 1/m^2)\,(7 - 3/m^2)\,\nu_{11}\,k^4/48$$

$$\mu_{42} = \nu_{42} - (1 - 1/n^2)\,\nu_{40}\,l^2/12 - (1 - 1/m^2)\,\nu_{22}\,k^2/2$$

$$+ (1 - 1/m^2)\,(1 - 1/n^2)\,\nu_{20}\,k^2l^2/24$$

$$+ (1 - 1/m^2)\,(7 - 3/m^2)\,\nu_{02}\,k^4/240 - (1 - 1/m^2)\,(7 - 3/m^2)\,(1 - 1/n^2)\,k^4l^2/2880$$

$$\mu_{33} = \nu_{33} - (1 - 1/m^2)\,\nu_{13}\,k^2/4 - (1 - 1/n^2)\,\nu_{31}\,l^2/4$$

$$+ (1 - 1/m^2)\,(1 - 1/n^2)\,\nu_{11}\,k^2l^2/16\,.$$

For m and n infinite these results give the formulas for two continuous variables already found by Baten[5] and Wold.[6]

The reader will note that this development does not impose the "high contact" condition, except in so far as it assumes the existence of the moments that occur in the formulas. And it exhibits in the clearest fashion that Sheppard's corrections are corrections on the average.

UNIVERSITY OF MICHIGAN.

[5] W. D. Baten: Corrections for the Moments of a Frequency Distribution in Two Variables; Annals of Mathematical Statistics, vol. II (1931), pp. 309–319.

[6] Loc. cit., p. 253.

FUNDAMENTALS OF THE THEORY OF INVERSE SAMPLING[1]

By Ching-Lai Shen

Part I. Introduction[2]

Section I. Statistical Concepts of the Theory of Sampling

One of the chief objects in statistics is to form a judgment of a very large statistical universe, known as a parent population, by means of a study of a part or sample thereof, which is drawn at random. To make a complete survey of the parent population is sometimes impossible or impractical. For example, it is impossible to measure the heights of all adult persons in a country. It is impractical to test for infectious bacteria the whole body of water in a city reservoir. All that we can do is to obtain an unbiased sample. By an unbiased sample, we mean a sample in which each individual has an equal and independent chance to be included. From this chosen sample we attempt to draw some conclusion concerning the nature of the whole parent population in accordance with certain mathematical principles.

Now the sample which we choose is of course only one of the samples that can be possibly drawn from a given parent population. Suppose there is a population of s individuals from which we wish to choose a sample of r. It is clear that there exist ${}_sC_r$ such samples, each of which is equally likely to be chosen. Therefore these ${}_sC_r$ samples constitute the so-called distribution of samples. To describe from the statistical point of view the distribution of samples, we must find its mean, standard deviation, skewness, excess, and other higher characteristics. The first three are usually referred to as elementary statistical functions.

Suppose x_i be the variate (by which we mean the magnitude of a specified character of an individual to be measured) where $i = 1, 2, 3, \cdots s$; and z_j be the samples chosen from the parent population where $j = 1, 2, 3, \cdots {}_sC_r$. Then the ${}_sC_r$ samples, each consisting of r variables, will be formed after the following fashion:

$$z_1 = x_1 + x_2 + x_3 + \cdots + x_r$$

$$z_2 = x_2 + x_3 + x_4 + \cdots + x_{r+1}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$z_{\binom{s}{r}} = x_{s-r+1} + x_{s-r+2} + x_{s-r+3} + \cdots + x_s$$

[1] A dissertation submitted in partial fulfillment of the requirement for the degree of doctor of philosophy in the University of Michigan.

[2] The writer wishes to express his appreciation for the assistance Professor H. C. Carver has given him in making this study.

If we denote the nth moment of the parent population about its mean by

$$\bar{u}_{n:x} = \frac{\sum_{i=1}^{s} (x_i - M_x)^n}{s}$$

and the nth moment of the distribution of samples about its mean by

$$\bar{\mu}_{n:z} = \frac{\sum_{j=1}^{\binom{s}{r}} (z_j - M_z)^n}{\binom{s}{r}}$$

and if we then utilize the multinomial theorem, we may be able to express the sample moments in terms of the moments of the parent population:[3]

$$(1) \quad \begin{cases} M_z = rM_x \\ \bar{\mu}_{2:z} = 2!\left\{P_2 \dfrac{s\bar{\mu}_{2:x}}{2!}\right\} \\ \bar{\mu}_{3:z} = 3!\left\{P_3 \dfrac{s\bar{\mu}_{3:x}}{3!}\right\} \\ \bar{\mu}_{4:z} = 4!\left\{P_4 \dfrac{s\bar{\mu}_{4:x}}{4!} + \dfrac{P_2^2}{2!} \dfrac{s^2\bar{\mu}_{2:x}^2}{(2!)^2}\right\} \\ \bar{\mu}_{5:z} = 5!\left\{P_5 \dfrac{s\bar{\mu}_{5:x}}{5!} + P_3P_2 \dfrac{s^2\bar{\mu}_{3:x}\bar{\mu}_{2:x}}{3!\,2!}\right\} \\ \bar{\mu}_{6:z} = 6!\left\{P_6 \dfrac{s\bar{\mu}_{6:x}}{6!} + P_4P_2 \dfrac{s^2\bar{\mu}_{4:x}\bar{\mu}_{2:x}}{4!\,2!}\right\} \\ \qquad + \dfrac{P_3^2}{2!} \dfrac{s^2\bar{\mu}_{3:x}^2}{(3!)^2} + \dfrac{P_2^3}{3!} \dfrac{s^3\bar{\mu}_{2:x}^3}{(2!)^3}, \text{ etc.} \end{cases}$$

where P_n is obtained from the sampling polynomial $P_n(\rho)$ by writing ρ^i as ρ_i:

$$(2) \quad \begin{cases} P_1(\rho) = \rho \\ P_2(\rho) = \rho - \rho^2 \\ P_3(\rho) = \rho - 3\rho^2 + 2\rho^3 \\ P_4(\rho) = \rho - 7\rho^2 + 12\rho^3 - 6\rho^4 \\ P_5(\rho) = \rho - 15\rho^2 + 50\rho^3 - 60\rho^4 + 24\rho^5 \\ P_6(\rho) = \rho - 31\rho^2 + 180\rho^3 - 390\rho^4 + 360\rho^5 - 120\rho^6, \text{ etc.} \end{cases}$$

where

$$\rho_i = \frac{r(r-1)(r-2)\cdots(r-\overline{i-1})}{s(s-1)(s-2)\cdots(s-\overline{i-1})}$$

[3] Carver, H. C., *Annals of Mathematical Statistics*, Vol. I, No. I, pp. 106–107.

Section II. Frequency Curve of the Distribution of Samples

The frequency distribution of samples is usually less scattered than individual observations. In order to ascertain the manner of the distribution, we have access to the well-known Type A Curve of Charlier.[4]

$$F(t) = \phi(t) - \frac{c_3}{3!}\phi^{(3)}(t) + \frac{c_4}{4!}\phi^{(4)}(t) - \frac{c_5}{5!}\phi^{(5)}(t) + \cdots \tag{3}$$

where $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$

$$\begin{aligned} c_3 &= \alpha_3 \\ c_4 &= \alpha_4 - 3 \\ c_5 &= \alpha_5 - 10\alpha_3 \\ c_6 &= \alpha_6 - 15\alpha_4 + 30 \\ c_7 &= \alpha_7 - 21\alpha_5 + 10\alpha_3 \\ c_8 &= \alpha_8 - 28\alpha_6 + 210\alpha_4 - 315, \text{ etc.} \end{aligned}$$

This formula is a powerful tool for representing any frequency; but it is emphasized by more than one author[5] that the usefulness of such a series representation of a frequency distribution depends upon the rapidity of convergence, and the rapidity of convergence in turn depends upon the extent to which the function $\phi(t)$ is a fair approximation for $F(t)$. We shall not, however, discuss here the question of convergence. What we are interested in is to apply this series representation to the distribution of samples and see whether our numerical experimentation justifies the use of it.

TABLE I

Heights of 1000 Freshman Students

(Original Measurements Made to Nearest 0.1 in.)

Class	Frequency
58.5–60.4	2
60.5–62.4	13
62.5–64.4	76
64.5–66.4	167
66.5–68.4	335
68.5–70.4	264
70.5–72.4	106
72.5–74.4	29
74.5–76.4	7
76.5–78.4	1

[4] Camp, B. H., *The Mathematical Part of Elementary Statistics*, p. 226.

[5] Rietz, H. L., *Mathematical Statistics* p. 62.
Carver, H. C., Frequency Curves, *Handbook of Mathematical Statistics*, p. 115.

First of all, therefore, we take for our numerical example the heights of 1000 freshman students in the University of Michigan, as recorded in Table I, which are assumed to constitute our parent population.

From the above data we compute the first 6 moments as follows:

$$M_x = 67.91$$

$\bar{\mu}_{2:x}$	= 6.279,068	σ_x	= 2.505,81
$\bar{\mu}_{3:x}$	= 0.489,552	$\alpha_{3:x}$	= 0.031,11
$\bar{\mu}_{4:x}$	= 132.685,214	$\alpha_{4:x}$	= 3.365,36
$\bar{\mu}_{5:x}$	= 78.435,794	$\alpha_{5:x}$	= 0.793,92
$\bar{\mu}_{6:x}$	= 4574.080,554	$\alpha_{6:x}$	= 18.476,43

Now suppose from this parent population in which $s = 1000$, we wish to choose ${}_{1000}C_{100}$ samples, each consisting of 100 individuals. To characterize the distribution of these samples, we first make the following table:

TABLE II

Values of ρ_i and P_i for $s = 1000$, $r = 100$

ρ_1 =	.1
ρ_2 =	.009,909,909,91
ρ_3 =	.000,973,117,406
ρ_4 =	.000,094,676,417,6
ρ_5 =	.000,009,125,437,84
ρ_6 =	.000,000,871,272,959,5
P_1 =	.1
P_2 =	.090,090,090,09
P_3 =	.072,216,505,082
P_4 =	.041,739,980,994
P_5 =	−.005,454,352,918
P_6 =	−.065,789,272,230
P_2^2 =	.008,058,351,516
P_2P_3 =	.006,472,571,500
P_2P_4 =	.003,764,792,358
P_2^3 =	.000,715,593,194
P_3^2 =	.005,195,978,741

Substituting into formulae (1), we obtain the first six moments of the distribution of samples:

$$M_s = 6791$$

$\bar{\mu}_{2:s}$	= 565.621,622	σ_s	= 23.782,8
$\bar{\mu}_{3:s}$	= 35.353,734	$\alpha_{3:s}$	= .002,628
$\bar{\mu}_{4:s}$	= 958,720.852,854	$\alpha_{4:s}$	= 2.996,679
$\bar{\mu}_{5:s}$	= 198,538.702,142	$\alpha_{5:s}$	= .026,093
$\bar{\mu}_{6:s}$	= 2,704,514,780.791,465	$\alpha_{6:s}$	= 14.945,539

The coefficients of Charlier's Type A Curve turn out to be very small and rapidly decreasing:

$$\frac{c_3}{3!} = .000,438$$

$$\frac{c_4}{4!} = -.000,138$$

$$\frac{c_5}{5!} = -.000,016$$

$$\frac{c_6}{6!} = -.000,006$$

We therefore may be justified in considering this series representation of the sample distribution as converging rapidly to the normal curve. It may be interesting to note that even from a parent population which is very skew, the distribution of samples is nearly normal—as the following example will show:

TABLE III

Weights of 1000 Freshman Students

(Original Measurements Made to Nearest Pound)

Class	Frequency
85–	1
95–	8
105–	45
115–	132
125–	232
135–	244
145–	161
155–	97
165–	50
175–	16
185–	7
195–	3
205–	4

$M_x = 139.32$

$\bar{\mu}_{2:x} = 296.8343$ $\quad \sigma_x = 17.228,87$

$\bar{\mu}_{3:x} = 3,230.802$ $\quad \alpha_{3:x} = 0.631,74$

$\bar{\mu}_{4:x} = .351,180.14$ $\quad \alpha_{4:x} = 3.985,67$

$\bar{\mu}_{5:x} = 11,811,480.5$ $\quad \alpha_{5:x} = 7.780,71$

$\bar{\mu}_{6:x} = 886,585,271$ $\quad \alpha_{6:x} = 33.898,36$

$$M_z = 13{,}932$$
$$\mu_{2:z} = 26{,}741.828{,}829 \qquad \sigma_z = 163.529$$
$$\mu_{3:z} = 233{,}317.229{,}045 \qquad \alpha_{3:z} = .05334$$
$$\mu_{4:z} = 2{,}144{,}736{,}851.477{,}805 \qquad \alpha_{4:z} = 2.9991$$
$$\mu_{5:z} = 62{,}008{,}368{,}279.121{,}883 \qquad \alpha_{5:z} = .53024$$
$$\mu_{6:z} = 287{,}107{,}828{,}746{,}809.017 \qquad \alpha_{6:z} = 15.00633$$

$$\frac{c_3}{3!} = .008{,}89$$

$$\frac{c_4}{4!} = -.000{,}04$$

$$\frac{c_5}{5!} = -.000{,}03$$

$$\frac{c_6}{6!} = .000{,}03$$

Indeed the distribution of samples, in general, is very nearly normal irrespective of the law of distribution of the parent population. From the practical point of view, as Professor H. C. Carver has remarked, the parent population has little control over the shape of the distribution of the samples of r is fifty or greater and if S is at least ten times as large as r.[6]

Now as a numerical illustration of the theory of sampling I may, for example, choose at random 100 weights from the parent population of 1000 weights of freshman students, as recorded in Table III, with the aim of ascertaining the probability that the mean of this sample exceeds 142 pounds.

Since we define the mean of a sample simply as the average measurement of the r individuals in the sample, which in this case is 100, it therefore follows that the ordinary moments of the distribution of sample means differ from those of the distribution of samples in (1) only by a constant multiple of $1/r^k$ where k is the order of the moments concerned, while the standardized moments remain unchanged. Therefore in this problem, we have the mean of the sample means equal to 139.32 and the standard deviation equal to 1.63529. The average weight, 142 pounds, may be expressed in standard units as

$$t = \frac{z - M_z}{\sigma_z} = \frac{142 - 139.32}{1.63529} = 1.63885$$

In accordance with (3), the probability that the mean of the sample exceeds 142 pounds is therefore equal to

$$P = \int_{1.63885}^{\infty} \left[\phi(t) - \frac{c_3}{3!}\phi^{(3)}(t) + \frac{c_4}{4!}\phi^{(4)}(t) - \frac{c_5}{5!}\phi^{(5)}(t) + \cdots\right] dt$$

[6] Carver, H. C., *Annals of Mathematical Statistics*, Vol. I, No. I, p. 112.

If we take the first term only, $P = \int_{1.63885}^{\infty} \phi(t)dt = .05062.$

If we take the first two terms, $P = \int_{1.63885}^{\infty} \phi(t)dt - .00889\phi^{(2)}(t)\Big]_{1.63885}^{\infty} = .05218.$

If we take the first three terms,

$$P = \int_{1.63885}^{\infty} \phi(t)dt - .00889\phi^{(2)}(t)\Big]_{1.63885}^{\infty} + (-.000,04)\phi^{(3)}(t)\Big]_{1.63885}^{\infty} = .052182.$$

Section III. Pearsonian Types of Curves

Charlier's Type A Series is, however, not the only known analytic representation of a frequency distribution. There are Pearsonian Types of Curves, the characteristics of which I shall need to summarize briefly. These Pearsonian Types of Curves are essential to the later development of our theory.

The curves, suggested by certain geometrical properties of unimodal frequency distribution, are all obtained from the solution of the differential equation:

$$\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{f(t)}$$

where $f(t)$ is assumed to be possibly expanded into a convergent power series, that is, $f(t) = b_0 + b_1t + b_2t^2 + \cdots$. When the first three terms of the power series are taken, the differential equation immediately takes the form of $\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{b_0 + b_1t + b_2t^2}$. The parameters, a, b_0, b_1, b_2, may be expressed in terms of moments:[7]

$$a = -\frac{\alpha_3}{2(1 + 2\delta)} \qquad b_0 = \frac{2 + \delta}{2(1 + 2\delta)}$$

$$b_1 = \frac{\alpha_3}{2(1 + 2\delta)} \qquad b_2 = \frac{\delta}{2(1 + 2\delta)}$$

where

$$\delta = \frac{2\alpha_4 - 3\alpha_3^2 - 6}{\alpha_4 + 3}$$

Based upon the difference in the nature of the roots of the equation $b_0 + b_1t + b_2t^2 = 0$, there have been derived thirteen types or curves. Of the particularly noteworthy ones, the normal curve and Type III may be mentioned. The criterion for the normal curve is $\alpha_3 = \delta = 0$; that for Type III is

[7] Carver, H. C., Frequency Curves, *Handbook of Mathematical Statistics*, p. 104.

$\delta = 0$ and $\alpha_3 \neq 0$. In order to fix the form in a particular case, we may refer to Pearson's Chart $\beta_1\beta_2$ Distribution[8] where

$$\beta_1 = \frac{\bar{\mu}_3^2}{\bar{\mu}_2^3} = \alpha_3^2, \qquad \beta_2 = \frac{\bar{\mu}_4}{\bar{\mu}_2^2} = \alpha_4,$$

and

$$K = \frac{b_1^2}{4b_0 b_2} = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)} = \frac{\alpha_3^2}{4\delta(2 + \delta)},$$

or to Elderton's *Frequency Curves and Correlation.*[9]

Section IV. The Inverse Sampling, Our Problem

It is now our problem to study the theory of inverse sampling, by which we mean that given the characteristics of a single sample drawn at random from a parent population, we wish to ascertain the probability that the corresponding characteristics of that parent population do not differ from those observed in the sample by more than a specified amount. To illustrate, suppose we are interested in knowing the average height of 1000 freshman students to which reference has already been made. Due to the fact that it takes too much time or is otherwise impractical to measure all of them so as to obtain the true average, we select at random one hundred of them and measure the heights of these one hundred individuals. Suppose the mean, the standard deviation, and the skewness of this sample of one hundred are computed and they are as follows:

$$M = \quad 67.99$$

$$\sigma = \quad 2.327$$

$$\alpha_3 = - \quad .12299$$

Now assuming that the true mean of the entire 1000 heights is unknown, let us find the probability that the true mean of this parent population lies between $M_x = a$ and $M_x = b$ by what we know of the characteristics of the observed sample of one hundred as recorded above. It is clear that if we can obtain an equation, $y = f(M_x)$, of the frequency curve associated with the distribution of hypothetical means of this parent population, we shall be able to ascertain the probability we desire by evaluating the following integral expression:

$$P = \frac{\int_a^b f(M_x)dM_x}{\int_{-\infty}^{\infty} f(M_x)dM_x}$$

---

[8] Pearson, K., *Tables for Statisticians and Biometricians*, Vol. II, front page.

[9] Elderton, W. P., *Frequency Curves and Correlation*, Table VI, opposite p. 46.

In the same way we can find the probability that the standard deviation of the parent population lies between two definite limits or that the skewness of the parent population lies between two definite limits.

Our procedure will therefore be as follows: First, assuming the *a priori* existence of a continuous sequence of hypothetical means of the parent population, we investigate the relation between the distribution of these hypothetical means of the parent population and the distribution of sample means. If such a relation exists, we shall be able to find an expression for the most probable value of the parent mean. Assuming the most probable value of the parent mean to be the true mean of the parent population, we shall obtain an expression for the most probable value of the standard deviation of the parent population. Then it will be possible for us to express the frequency curve associated with the distribution of hypothetical means of the parent population in the form of $f(M_x)$. Similarly we may find the frequency functions associated with the standard deviation and skewness of the parent population.

Before leaving this section, it is perhaps not out of place to say a word about the connection of this theory of inverse sampling with Bayes's Theorem. The theory of inverse sampling (which deals essentially with the problem of judging the nature of a whole by observation of a part of it) belongs to the domain of inductive probability, or inverse probability, upon which Bayes's Theorem was founded. In order to solve a problem of inductive probability, it is necessary to postulate the *a priori* existence of the causes from which an event takes place, which, in our case, is the hypothetical means of the parent population.

This *a priori* hypothesis which gives rise to Bayes's Theorem has been viewed with suspicion by a number of mathematical statisticians. For example, the theorem has been called into question by such mathematicians as Bing, Venn, Chrystal, and others, including several now living. But so far as the present writer is aware, no definite conclusion has been reached. It is true that on the one hand Bayes's Theorem has not been rigidly demonstrated and proved by logic; but on the other hand the process of generalization from observational data is justified within the limits of ordinary practical application. One who holds Bayes's Theorem strongly may even say that the *a priori* hypothesis is absolutely necessary to scientific inferences. Concerning this controversy, Pearson takes a liberal point of view: "I hold this theorem [Bayes's Theorem] not as rigidly demonstrated, but I think with Edgeworth that the hypothesis of the equal distribution of ignorance is within the limits of practical life justified by experience of statistical ratios, which *a priori* are unknown · · · ."[10] He has further remarked that "the practical man · · · will accept the results of inverse probability of Bayes-Laplace brand till better are forthcoming."[11] Using

[10] Pearson, K., On the Influence of Past Experience on Future Expectation, *Philosophical Magazine*, Vol. 13, Jan.–June, 1907, p. 366.

[11] Pearson, K., The Fundamental Problem of Practical Statistics, *Biometrika*, Vol. 13, 1920–21, p. 3.

Pearson's viewpoint, we shall proceed with our problem by postulating *a priori* the existence of hypothetical means of the parent population from which our sample is drawn.

Part II. Fundamental Relation between the Moments of the Distribution of Sampling Means and the Moments of the Distribution of the Hypothetical Means Associated with the Parent Population

The characteristics of the distribution of sample means, as we have pointed out in Part I, Section II, differ from those of the sample distribution only by a constant multiple of $(1/r)^k$ where k is the order of the moments concerned. We may write down the first six moments of the distribution of sample means:

$$
(4)\quad
\begin{cases}
M_{z_x} = M_x \\
\bar{\mu}_{2:z_x} = 2!\,\dfrac{s}{r^2}\left\{P_2\,\dfrac{\bar{\mu}_{2:x}}{2!}\right\} \\
\bar{\mu}_{3:z_x} = 3!\,\dfrac{s}{r^3}\left\{P_3\,\dfrac{\bar{\mu}_{3:x}}{3!}\right\} \\
\bar{\mu}_{4:z_x} = 4!\,\dfrac{s}{r^4}\left\{P_4\,\dfrac{\bar{\mu}_{4:x}}{4!} + \dfrac{P_2^2}{2!}\,\dfrac{s\,\bar{\mu}_{2:x}^2}{(2!)^2}\right\} \\
\bar{\mu}_{5:z_x} = 5!\,\dfrac{s}{r^5}\left\{P_5\,\dfrac{\bar{\mu}_{5:x}}{5!} + P_3P_2\,\dfrac{s\,\bar{\mu}_{3:x}\,\bar{\mu}_{2:x}}{3!\,2!}\right\} \\
\bar{\mu}_{6:z_x} = 6!\,\dfrac{s}{r^6}\left\{P_6\,\dfrac{\bar{u}_{6.x}}{6!} + P_4P_2\,\dfrac{s\,\bar{\mu}_{4:x}\,\bar{\mu}_{2:x}}{4!\,2!} + \dfrac{P_3^2}{2!}\,\dfrac{s\,\bar{\mu}_{3:x}^2}{(3!)^2} + \dfrac{P_2^3}{3!}\,\dfrac{s^2\,\bar{\mu}_{2:x}^3}{(2!)^3}\right\}
\end{cases}
$$

From these we immediately obtain

$$
(5)\quad
\begin{cases}
M_{z_x} = M_x \\
\sigma_{z_x} = \sqrt{\dfrac{s-r}{r(s-1)}}\,\sigma_x \\
\alpha_{3:z_x} = \dfrac{s-2r}{s-2}\sqrt{\dfrac{s-1}{r(s-r)}}\,\alpha_{3:x} \\
\alpha_{4:z_x} - 3 = \dfrac{(s-1)(s^2+s-6rs+6r^2)}{r(s-r)(s-2)(s-3)}\{\alpha_{4:x}-3\} \\
\qquad\qquad - \dfrac{6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)},\ \text{etc.}
\end{cases}
$$

If our parent population is infinite, which is a special case by allowing $s \to \infty$, then we have

$$(6)\qquad \begin{cases} M_{z_x} = M_x \\ \sigma_{z_x} = \dfrac{1}{\sqrt{r}}\,\sigma_{3:x} \\ \alpha_{3:z_x} = \dfrac{1}{\sqrt{r}}\,\alpha_{3:x} \\ \alpha_{4:z_x} - 3 = \dfrac{1}{r}\,(\alpha_{4:x} - 3), \text{ etc.} \end{cases}$$

Let us now define $f(t)$ as a frequency function of the distribution of sample means z_x in standard units, i.e.,

$$(7)\qquad t = \frac{Z_x - M_{z_x}}{\sigma_{z_x}}$$

Denoting the observed mean of a given sample by m_1 and making proper substitutions of (5), we obtain

$$(8)\qquad t = \frac{m_1 - M_x}{\sigma_{z_x}} = \frac{m_1 - M_x}{\sqrt{\dfrac{s - r}{r(s - 1)}}\,\sigma_x}$$

It is clear that if we hold s, r and σ_x constant and let M_x vary, then t is a function of M_x only and consequently $f(t)$ becomes a function of M_x.

Suppose now $M_x^{(1)}$, $M_x^{(2)}$, $M_x^{(3)}$, $\cdots$ be a continuous sequence of hypothetical means, which M_x has an equal chance to assume. These hypothetical means will certainly lie in a linear interval between their natural limits. Then the probability that M_x lies between $M_x \pm \frac{1}{2}\,dM_x$ is $f(t)dM_x$. Therefore, to obtain the probability that M_x lies in the interval $M_x^{(i)} \leqq M_x \leqq M_x^{(i+1)}$, it is only necessary to carry out the integration of this expression:

$$(9)\qquad \int_{M_x^{(i)}}^{M_x^{(i+1)}} f(t)\,dM_x$$

There is no question as regards the existence of this integral in case of an infinite parent population. As for a finite population, we may still use this continuous function as an interpolation function to the true discontinuous function.

Let us now define $P(t)$ as the probability function for which the hypothetical mean of the parent population falls within certain specified limits. Considering

$\mu_{n:p}$ as the nth moment of this probability function about a fixed point, we will have the following relation:

$$\mu_{n:p} = \frac{\int_{-l}^{l} M_x^n f(t)dM_x}{\int_{-l}^{l} f(t)dM_x} \tag{10}$$

where l and $-l$ are their natural limits.

Since from (8), $M_x = m_1 - \sigma_{z_x}t$, then after substitution, we obtain

$$\mu_{n:p} = \frac{\sigma_{z_x}\int_{\frac{m_1-l}{\sigma_{z_x}}}^{\frac{m_1+l}{\sigma_{z_x}}} (m_1 - \sigma_{z_x}t)^n f(t)dt}{\sigma_{z_x}\int_{\frac{m_1-l}{\sigma_{z_x}}}^{\frac{m_1+l}{\sigma_{z_x}}} f(t)dt}$$

$$= \int_{\frac{m_1-l}{\sigma_{z_x}}}^{\frac{m_1+l}{\sigma_{z_x}}} (m_1 - \sigma_{z_x}t)^n f(t)dt \tag{11}$$

$$= m_1^n - \binom{n}{1} m_1^{n-1}\bar{\mu}_{1:z_x} + \binom{n}{2} m_1^{n-2}\bar{\mu}_{2:z_x} - \binom{n}{3} m_1^{n-3}\bar{\mu}_{3:z_x} + \cdots + (-1)^n\binom{n}{n}\bar{\mu}_{n:z_x}$$

$$\begin{cases} \mu_{1:p} = M_p = m_1 \\ \mu_{2:p} = m_1^2 + \bar{\mu}_{2:z_x} \\ \mu_{3p} = m_1^3 + 3m_1\bar{\mu}_{2:z_x} - \bar{\mu}_{3:z_x} \\ \mu_{4:p} = m_1^4 + 6m_1^2\bar{\mu}_{2:z_x} - 4m_1\bar{\mu}_{3:z_x} + \bar{\mu}_{4:z_x}, \text{ etc.} \end{cases} \tag{12}$$

The first relation $M_p = m_1$ is important because it shows that the mean of the hypothetical means of the parent population is equal to the mean of the observed sample drawn from it. To state this in a theorem, we will have

Theorem I. The expected value of a parent mean is equal to the mean of an observed sample chosen from the parent population.

We now wish to express the moments of the probability function about its mean in terms of the moments of sample distribution. In general, the nth moment of any frequency distribution about its mean, $\bar{\mu}_n$, can be expressed in terms of its moments about a fixed point after the following fashion:

$$\bar{\mu}_n = \mu_n - \binom{n}{1} M\mu_{n-1} + \binom{n}{2} M^2\mu_{n-2} - \cdots + (-1)^n\binom{n}{n} M^n. \tag{13}$$

Therefore when we substitute (11) into (13) we obtain

$$
\begin{aligned}
\bar{\mu}_{n:p} = m_1^n &- \binom{n}{1} m_1^{n-1}\, \bar{\mu}_{1:s_x} + \binom{n}{2} m_1^{n-2}\, \bar{\mu}_{2:s_x} - \cdots \\
&+ (-1)^{n-3}\binom{n}{n-3} m_1^3\, \bar{\mu}_{n-3:s_x} + (-1)^{n-2}\binom{n}{n-2} m_1^2\, \bar{\mu}_{n-2:} \\
&+ (-1)^{n-1}\binom{n}{n-1} m_1\, \bar{\mu}_{n-1:s_x} + (-1)^n \binom{n}{n} \bar{\mu}_{n:s_x} \\
- m_1 \binom{n}{1} \Bigg[m_1^{n-1} &- \binom{n-1}{1} m_1^{n-2}\, \bar{\mu}_{1:s_x} + \binom{n-1}{2} m_1^{n-3}\, \bar{\mu}_{2:s_x} \\
&+ (-1)^{n-3}\binom{n-1}{n-3} m_1^2\, \bar{\mu}_{n-3:s_x} + (-1)^{n-2}\binom{n-1}{n-2} m_1\, \bar{\mu}_{n-2:} \\
&+ (-1)^{n-1}\binom{n-1}{n-1} \bar{\mu}_{n-1:s_x} \Bigg] \\
+ m_1^2 \binom{n}{2} \Bigg[m_1^{n-2} &- \binom{n-2}{1} m_1^{n-3}\, \bar{\mu}_{1:s_x} + \binom{n-2}{2} m_1^{n-4}\, \bar{\mu}_{2:s_x} - \cdots \\
&+ (-1)^{n-3}\binom{n-2}{n-3} m_1\, \bar{\mu}_{n-3:s_x} + (-1)^{n-2}\binom{n-2}{n-2} \bar{\mu}_{n-2:s_x} \Bigg] \\
- m_1^3 \binom{n}{3} \Bigg[m_1^{n-3} &- \binom{n-3}{1} m_1^{n-4}\, \bar{\mu}_{1:s_x} + \binom{n-3}{2} m_1^{n-5}\, \bar{\mu}_{2:s_x} - \cdots \\
&+ (-1)^{n-3}\binom{n-3}{n-3} \bar{\mu}_{n-3:s_x} \Bigg] \\
+ \cdots\cdots&\cdots\cdots\cdots\cdots\cdots\cdots \\
&+ (-1)^{n-1} m_1^{n-1} \binom{n}{n-1} \left[m_1 - \binom{1}{1} \bar{\mu}_{1:s_x} \right] \\
&+ (-1)^n m_1^n \binom{n}{n} .
\end{aligned}
$$

Adding vertically each column, we obtain

$$
\begin{aligned}
\bar{\mu}_{x:p} = m_1^n &\left[\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \binom{n}{4} - \cdots + (-1)^n \binom{n}{n} \right] \\
- m_1^{n-1}\, \bar{\mu}_{1:s_x} &\Bigg[\binom{n}{1}\binom{n}{0} - \binom{n-1}{1}\binom{n}{1} + \binom{n-2}{1}\binom{n}{2} - \binom{n-3}{1}\binom{n}{3} \\
&+ \binom{n-4}{1}\binom{n}{4} - \cdots + (-1)^{n-1}\binom{n}{n-1} \Bigg]
\end{aligned}
$$

$$+ m_1^{n-2}\bar{\mu}_{2:sx}\left[\binom{n}{2}\binom{n}{0} - \binom{n-1}{2}\binom{n}{1} + \binom{n-2}{2}\binom{n}{2} - \binom{n-3}{2}\binom{n}{3} + \binom{n-4}{2}\binom{n}{4} - \cdots + (-1)^{n-2}\binom{n}{n-2}\right]$$

$$-\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$+ (-1)^{n-1} m_1 \bar{\mu}_{n-1:sx}\left[\binom{n}{n-1}\binom{n}{0} - \binom{n-1}{n-1}\binom{n}{1}\right] + (-1)^n \bar{\mu}_{n:sx}$$

The first row of the above expression is equal to $m_1^n(1 - 1)^n = 0$; the second row is equal to

$$- m_1^{n-1}\bar{\mu}_{1:sx}\left[\frac{n!}{0!\,n!}\cdot\frac{n!}{1!(n-1)!} - \frac{n!}{1!\,(n-1)!}\cdot\frac{(n-1)!}{1!\,(n-2)!} + \frac{n!}{2!\,(n-2)!}\,\frac{(n-2)!}{1!\,(n-3)!} - \cdots + (-1)^n\frac{n!}{(n-1)!\,1!}\right]$$

$$= -m_1^{n-1}\bar{\mu}_{1:sx}\frac{n!}{1!}\left[\frac{1}{0!\,(n-1)!} - \frac{1}{1!\,(n-2)!} + \frac{1}{2!\,(n-3)!} - \cdots + (-1)^n\frac{1}{(n-1)!\,0!}\right]$$

$$= -m_1^{n-1}\bar{\mu}_{1:sx}\frac{n}{1!}[1 - {}_{n-1}C_1 + {}_{n-1}C_2 - \cdots + (-1)^n\,{}_{n-1}C_{n-1}]$$

$$= -m_1^{n-1}\bar{\mu}_{1:sx}\frac{n}{1!}(1 - 1)^{n-1} = 0;$$

the third row is equal to

$$m_1^{n-2}\bar{\mu}_{2:sx}\left[\frac{n!}{0!\,n!}\cdot\frac{n!}{2!\,(n-2)!} - \frac{n!}{1!\,(n-1)!}\cdot\frac{(n-1)!}{2!\,(n-3)!} + \frac{n!}{2!\,(n-2)!}\cdot\frac{(n-2)!}{2!\,(n-4)!} - \cdots + (-1)^{n-1}\frac{n!}{(n-2)!\,2!}\right]$$

$$= m_1^{n-2}\bar{\mu}_{2:sx}\frac{n(n-1)}{2!}[1 - {}_{n-2}C_1 + {}_{n-2}C_2 - \cdots + (-1)^{n-1}\,{}_{n-2}C_{n-2}]$$

$$= m_1^{n-2}\bar{\mu}_{2:sx}\frac{n(n-1)}{2!}(1 - 1)^{n-2} = 0;$$

and similarly all the other rows turn out to be zero except the last one which is equal to $(-1)^n\bar{\mu}_{n:sx}$

(14) $$\bar{\mu}_{n:p} = (-1)^n\bar{\mu}_{n:sx}$$

This may be rewritten as

$$(15) \qquad \begin{cases} \bar{\mu}_{2n:p} = \bar{\mu}_{2n:\bar{x}} \\ \bar{\mu}_{2n+1:p} = -\bar{\mu}_{2n+1:\bar{x}} \end{cases}$$

or in standard units

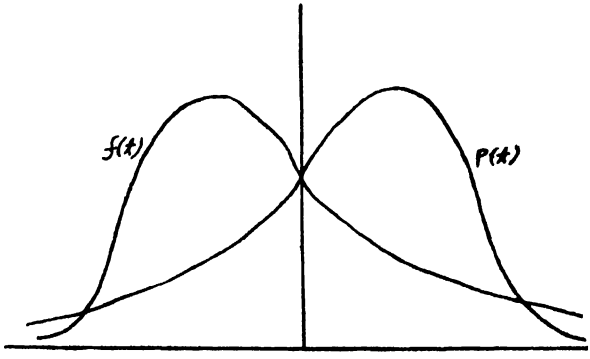
$$\begin{cases} \alpha_{2n:p} = \alpha_{2n:\bar{x}} \\ \alpha_{2n+1:p} = -\alpha_{2n+1:\bar{x}} \end{cases}$$

The results[12] of (15) are important and fundamental because they establish the relation between the Theory of Inverse Sampling and the Theory of Sampling. Therefore we may formulate the following theorems:

Theorem II. The even moments of the distribution of the hypothetical means of a parent population about its mean are equal to the corresponding even moments of the distribution of the sample means about the mean.

Theorem III. The odd moments of the distribution of the hypothetical means of a parent population about its mean are equal to the negative of the corresponding odd moments of the distribution of the sample means about the mean.

Since the even moments of the two distributions are the same, while the odd moments differ only in sign, it is evident that for symmetrical distributions, the two curves $f(t)$ and $P(t)$ are exactly identical, because in a symmetrical distribution all the odd moments about the mean are bound to vanish. In case of nonsymmetrical distributions, the curve $P(t)$ is nothing but a vertical reflection of the curve $f(t)$ as shown in the figure:



In other words, if $f(t)$, for instance, assumes Pearson's Type III Function, then $P(t)$ also assumes Pearson's Type III Function except that their skewness is different in sign though equal numerically. We therefore state our theorem as follows:

[12] So far as the writer is aware, these theorems were first developed by Professor H. C. Carver.

Theorem IV. The curves for the distribution of the hypothetical means of the parent population and the curve for the distribution of the means of the sample obtained from the parent population are symmetrically situated and one is a vertical reflection of the other.

Part III. Inverse Sampling Associated with a Normal Parent Population

We shall be concerned in this part of our discussion with a normal parent population. In accordance with the characteristics of a normal parent population we wish to investigate the most probable values of its mean and variance, thereby obtaining the distributions of the hypothetical means and variances of the parent population.

Section I. Most Probable Value of the Mean of the Parent Population

In Part I, Section III, we have mentioned Pearsonian Types of Frequency Curves whose differential equation is

$$\frac{1}{t}\frac{dy}{dt} = \frac{a - t}{b_0 + b_1 t + b_2 t^2}.$$

It is clear that the mode of these curves is at $t = a$, provided the mode exists. But to recapitulate:

$$a = \frac{-\alpha_3}{2(1 + 2\delta)},$$

where

$$\delta = \frac{2\alpha_4 - 3\alpha_3^2 - 6}{\alpha_4 + 3};$$

consequently for the mode of the distribution of sample means, we have

$$t = \frac{-\alpha_{3:z_x}}{2(1 + 2\delta_{z_x})}, \tag{16}$$

where

$$\delta_{z_x} = \frac{2\alpha_{4:z_x} - 3\alpha_{3:z_x}^2 - 6}{\alpha_{4:z_x} + 3} \tag{17}$$

$$= \frac{\begin{array}{r}2(s-1)(s-2)(s^2+s-6rs+6r^2)(\alpha_{4:x}-3) \\ -12s(r-1)(s-2)(s-r-1) - 3(s-1)(s-3)(s-2r)^2\alpha_{3:x}^2\end{array}}{\begin{array}{r}(s-2)\{(s-1)(s^2+s-6rs+6r^2)(\alpha_{4:x}-3) \\ -6s(r-1)(s-r-1) + 6r(s-r)(s-2)(s-3)\}\end{array}}$$

$$\therefore t = \frac{z_x - M_{z_x}}{\sigma_{z_x}} = \frac{z_x - M_x}{\sqrt{\frac{s-r}{r(s-1)}}\,\sigma_x} = \frac{-\frac{s-2r}{s-2}\sqrt{\frac{s-1}{r(s-r)}}\,\alpha_{3:x}}{2(1 + 2\delta_{z_x})} \tag{18}$$

Now according to Theorem IV, the mode of the probability function $P(t)$ is situated symmetrically with respect to the mode of the frequency function $f(t)$ of the distribution of sample means; hence, for the mode of the probability function of hypothetical means of the parent population, we have

$$t = \frac{m_1 - M_x}{\sqrt{\frac{s-r}{r(s-1)}}\,\sigma_x} = \frac{\frac{s-2r}{s-2}\sqrt{\frac{s-1}{r(s-r)}}\,\alpha_{3:x}}{2(1+2\delta_{z_x})} \tag{19}$$

where δ_{z_x} remains unchanged because it is a function of $\alpha^2_{3:z_x}$ and $\alpha_{4:z_x}$, each being always positive.

Solving for M_x, which will now be the most probable value of the mean of the parent population and hence denoted by $\hat{M}_x$, we have

$$\hat{M}_x = m_1 - \frac{s-2r}{r(s-2)}\,\frac{\sigma_x \alpha_{3:x}}{2(1+2\delta_{z_x})} \tag{20}$$

It is interesting to note that if $s = 2r$, this expression yields $\hat{M}_x = m_1$, irrespective of the law of distribution of the parent population provided only that δ_{z_x} is not exactly equal to $-\frac{1}{2}$. But since the Pearson's function is used for graduation, one should not fail to see that the mode so obtained gives only an approximation to the true mode. Therefore we state a theorem as follows:

Theorem V. If a sample is composed of one-half of the variates of the parent population from which the sample is chosen, then the best approximated 'most probable value' of the mean of the parent population is equal to the mean of the observed sample provided only that δ_{z_x} is not exactly equal to $-\frac{1}{2}$.

It is further observed that if $\alpha_{3:x} = 0$ but $\delta_{z_x} \neq -\frac{1}{2}$, then the expression (20) will likewise yield $\hat{M}_x = m_1$. But $\alpha_{3:x} = 0$ implies that the frequency curve of the parent population is symmetrical. Hence

Theorem VI. For any symmetrical curves associated with the distribution of the parent population, the best approximated 'most probable value' of the mean of the parent population is equal to the mean of the observed sample provided δ_{z_x} is not exactly equal to $-\frac{1}{2}$.

But we will investigate further the most probable value of the mean of a normal parent population, and we know that in a normal distribution the moments bear the following relation:[13]

$$\begin{cases} \alpha_{2n} = \dfrac{(2n)!}{2^n\, n!} \\ \alpha_{2n+1} = 0 \end{cases} \tag{21}$$

[13] Carver, H. C., Frequency Curves, *Handbook of Mathematical Statistics*, p. 97.

i.e., $\alpha_3 = 0$
$\alpha_4 = 3$
$\alpha_5 = 0$
$\alpha_6 = 15$
$\alpha_7 = 0$
$\alpha_8 = 105$

etc.

Consequently for a normal parent population the α_{z_x} function in (17) is immediately reduced to

$$\delta_{z_x} = \frac{2s(r-1)(s-r-1)}{s(r-1)(s-r-1) - r(s-r)(s-2)(s-3)} \tag{22}$$

Let us, first of all, investigate the possibility that this expression will be exactly equal to $-\frac{1}{2}$ for positive integral values of r and s.

Suppose we set

$$\frac{2s(r-1)(s-r-1)}{s(r-1)(s-r-1) - r(s-r)(s-2)(s-3)} = -\frac{1}{2}$$

and solve r in terms of s. Thus we obtain

$$r = \frac{s}{2} \pm \frac{\sqrt{s^2(s^2 - 10s + 6)^2 + 20s(s-1)(s^2 - 10s + 6)}}{2(s^2 - 10s + 6)} \tag{23}$$

If $s \geqq 10$, then the second term on the right side is positive. As it is absurd that r should be greater than s, therefore the positive sign of the double sign should not be taken. Then, as the second term is obviously greater than $\frac{s}{2}$, the right member will be negative. Since r cannot be negative, no positive integral values of r and s, for which $s \geqq r$, can satisfy (23). For $s < 10$, there are only nine positive integers; and direct substitution of each will tell us that only when $s = 1$, 2, or 3, r is a positive integer which is either 1 or 2. As these are trifle cases because a parent population can never be so small, we may safely say that for a normal parent population

$$\hat{M}_x = m_1 \tag{24}$$

Theorem VII. For a normal parent population, the best approximated 'most probable value' of the mean of the parent population is equal to the mean of the observed sample from it.

For an infinite parent population, i.e., $s \to \infty$ (20) yields on reduction

$$\hat{M}_x = m_1 - \frac{1}{r} \frac{\sigma_x \alpha_{3:x}}{2(1 + 2\delta_{z_x})} \tag{25}$$

where

$$\delta_{z_x} = \frac{2(\alpha_{4:x} - 3) - 3\alpha_{3:x}^2}{(\alpha_{4:x} - 3) - 6r} \text{ [(from 17)'}$$

Formula (25) yields immediately $\hat{M}_x = m_1$ if $\alpha_{3:x} = 0$ and $\delta_{s_x} \neq -\frac{1}{2}$. For a normal parent population $\delta_{s_x} = 0$. Hence Theorem VI and Theorem VII both hold for the infinite case.

Section II. Most Probable Value of the Standard Deviation of the Parent Population

To find the most probable value of the standard deviation of the parent population, we shall assume the mean of the parent population to be the best approximated 'most probable value' of the mean, which we have obtained in the preceding section. This assumption is necessary since we do not know the true mean of the parent population.

Now, to start with, we shall consider ${}_sC_r$ possible samples, each consisting of r variables. The second moment of each sample computed about the best approximated 'most probable value' of the mean of the parent population may be written as

$$z_1 = \frac{1}{r}\{(x_1 - m_1)^2 + (x_2 - m_1)^2 + (x_3 - m_1)^2 + \cdots + (x_r - m_1)^2\}$$

$$z_2 = \frac{1}{r}\{(x_2 - m_1)^2 + (x_3 - m_1)^2 + (x_4 - m_1)^2 + \cdots + (x_{r+1} - m_1)^2\}$$

. . . . . .

$$z_{\binom{s}{r}} = \frac{1}{r}\{(x_{s-r+1} - m_1)^2 + (x_{s-r+2} - m_1)^2 + (x_{s-r+3} - m_1)^2 + \cdots + (x_s - m_1)^2\}$$

If we write $(x_i - m_1)^2 = y_i$, it is clear that the above may be considered as a distribution of sample means drawn from a parent population $y_1, y_2, y_3 \cdots y_s$; and consequently

$$(26) \qquad \begin{cases} M_{z_y} = M_y \\ \sigma_{z_y} = \sigma_y \sqrt{\dfrac{s-r}{r(s-1)}} \\ \alpha_{3:z_y} = \dfrac{s-2r}{s-2}\sqrt{\dfrac{s-1}{r(s-r)}}\,\alpha_{3:y} \\ \alpha_{4:z_y} - 3 = \dfrac{(s-1)(s^2+s-6rs+6r^2)}{r(s-r)s-2)(s-3)}\{\alpha_{4:y} - 3\} \\ \qquad\qquad - \dfrac{6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)}. \end{cases}$$

Now the nth moment of y about a fixed point may be written as

$$\mu_{n:y} = \frac{1}{N}\sum y^n = \frac{1}{N}\sum (x - m_1)^{2n}$$

$$= \frac{1}{N}\sum \{(x - M_x) + (M_x - m_1)\}^{2n}$$

$$(27) \qquad = \bar{\mu}_{2n:x} + \binom{2n}{1}\bar{\mu}_{2n-1:x}(M_x - m_1)$$

$$+ \binom{2n}{2}\bar{\mu}_{2n-2:x}(M_x - m_1)^2 + \binom{2n}{3}\bar{\mu}_{2n-3:x}(M_x - m_1)^3$$

$$+ \binom{2n}{4}\bar{\mu}_{2n-4:x}(M_x - m_1)^4 + \cdots + (M_x - m_1)^{2n}.$$

On the assumption that our parent population is normally distributed and due to the fact that in a normally distributed function

$$\alpha_{2n} = \frac{(2n)!}{2^n\, n!} \quad \text{and} \quad \alpha_{2n+1} = 0 \qquad \text{[See (21)]},$$

the expression (27) immediately takes this form:

$$(28) \qquad \mu_{n:y} = \frac{2n!}{2^n \cdot n!}\sigma_x^{2n} + \binom{2n}{2}\frac{(2n-2)!}{2^{n-1}(n-1)!}(M_x - m_1)^2\sigma_x^{2n-2}$$

$$+ \binom{2n}{4}\frac{(2n-4)!}{2^{n-2}(n-2)!}(M_x - m_1)^4\sigma_x^{2n-4} + \cdots + (M_x - m_1)^{2n}.$$

Imposing the condition mentioned at the beginning of this section (i.e., M_x assumes its best approximated 'most probable value' m_1), then all the terms drop out except the first one. Hence, as a final form, we have

$$(29) \qquad \mu_{n:y} = \frac{2n!}{2^n \cdot n!}\sigma_x^{2n}$$

$$(30) \qquad \begin{cases} \mu_{1:y} = M_y = \sigma_x^2 \\ \mu_{2:y} = 3\sigma_x^4 \\ \mu_{3:y} = 15\sigma_x^6 \\ \mu_{4:y} = 105\sigma_x^8 \\ \qquad \text{etc.} \end{cases}$$

It follows that the kth moment of y about its mean will be

$$\bar{\mu}_{k:y} = \frac{\sum (y - M_y)^k}{N} = \mu_{k:y} - \binom{k}{1} \mu_{k-1:y} M_y + \binom{k}{2} \mu_{k-2:y} M_y^2 - \cdots + (-1)^k \binom{k}{k} M_y^k$$

$$= \frac{2k!}{2^k \cdot k!} \sigma_x^{2k} - \binom{k}{1} \frac{(2k-2)!}{2^{k-1}(k-1)!} \sigma_x^{2k} + \binom{k}{2} \frac{(2k-4)!}{2^{k-2}(k-2)!} \sigma_x^2 - \cdots + (-1)^k \sigma_x^{2k}$$

$$= \sigma_x^{2k} \left[\frac{2k!}{2^k \cdot k!} - \binom{k}{1} \frac{(2k-2)!}{2^{k-1}(k-1)!} + \binom{k}{2} \frac{(2k-4)}{2^{k-2}(k-2)!} - \cdots + (-1)^k \right] \tag{31}$$

$$= \sigma_x^{2k} \frac{2k!}{2^k \cdot k!} \left[1 - \frac{k}{1!(2k-1)} + \frac{k(k-1)}{2!(2k-1)(2k-3)} - \frac{k(k-1)(k-2)}{3!(2k-1)(2k-3)(2k-5)} + \cdots + (-1)^k \frac{k!}{k!(2k-1)(2k-3)(2k-5)\cdots(3)\cdot(1)} \right]$$

$$\begin{cases} \bar{\mu}_{1:y} = 0 \\ \bar{\mu}_{2:y} = 2\sigma_x^4 \\ \bar{\mu}_{3:y} = 8\sigma_x^6 \\ \bar{\mu}_{4:y} = 60\sigma_x^8 \\ \bar{\mu}_{5:y} = 544\sigma_x^{10} \\ \bar{\mu}_{6:y} = 6040\,\sigma_x^{12} \\ \quad \text{etc.} \end{cases} \tag{32}$$

And therefore we obtain

$$\begin{cases} \alpha_{3:y} = 2\sqrt{2} \\ \alpha_{4:y} = 15 \\ \alpha_{5:y} = 68\sqrt{2} \\ \alpha_{6:y} = 715 \\ \quad \text{etc.} \end{cases} \tag{33}$$

Making proper substitution of (30), (32), (33) into (26), we obtain

$$(34)\quad \begin{cases} M_{z_y} = \sigma_x^2 \\ \sigma_{z_y} = \sqrt{\dfrac{2(s-r)}{r(s-1)}}\,\sigma_x^2 \\ \alpha_{3:z_y} = \dfrac{2(s-2r)}{s-2}\sqrt{\dfrac{2(s-1)}{r(s-r)}} \\ \alpha_{4:z_y} - 3 = \dfrac{12(s-1)(s^2+s-6rs+6r^2) - 6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)} \end{cases}$$

For an infinite parent population, i.e., $s \to \infty$, we have

$$(35)\quad \begin{cases} M_{z_y} = \sigma_x^2 \\ \sigma_{z_y} = \sqrt{\dfrac{2}{r}}\,\sigma_x^2 \\ \alpha_{3:z_y} = 2\sqrt{\dfrac{2}{r}} \\ \alpha_{4:z_y} - 3 = \dfrac{12}{r} \end{cases}$$

Now again with reference to Pearsonian Types of Curves for which the mode is at $t = a$, we have for the mode of the distribution of sample means z_y,

$$(36)\quad t = \frac{z_y - M_{z_y}}{\sigma_{z_y}} = -\frac{\alpha_{3:z_y}}{2(1+2\delta_{z_y})} \text{ where}$$

$$\delta_{z_y} = \frac{2\alpha_{4:z_y} - 3\alpha_{3:z_y}^2 - 6}{\alpha_{4:z_y} + 3}$$

$$(37)\quad = 2 - \frac{(s-3)[4(s-2r)^2(s-1) + 2r(s-2)^2(s-r)]}{(s-2)[2(s-1)(s^2+s-6rs+6r^2) + r(s-r)(s-2)(s-3) - s(r-1)(s-r-1)}$$

Substituting (34) into (36), we obtain

$$(38)\quad \frac{z_y - \sigma_x^2}{\sqrt{\dfrac{2(s-r)}{r(s-1)}}\,\sigma_x^2} = -\frac{s-2r}{s-2}\sqrt{\frac{2(s-1)}{r(s-r)}}\cdot\frac{1}{1+2\delta_{z_y}}$$

By Theorem IV, the best approximated 'most probable value' of the standard deviation of the parent population is obtained from (38) by changing the sign of the right member and replacing z_y by m_2. Thus we have

$$\frac{m_2 - \sigma_x^2}{\sqrt{\dfrac{2(s-r)}{r(s-1)}}\,\sigma_x^2} = \frac{s-2r}{s-2}\sqrt{\frac{2(s-1)}{r(s-r)}}\cdot\frac{1}{1+2\delta_{z_y}}$$

Solving for σ_x, which is now the best approximated 'most probable value' and should therefore be denoted by $\hat{\sigma}_x$, we then have

$$\hat{\sigma}_x^2 = \hat{\mu}_{2:x} = \frac{m_2}{1 + \dfrac{2(s-2r)}{r(s-2)(1+2\delta_{z_y})}} \tag{39}$$

The best approximate 'most probable value' of the standard deviation may therefore be written down as

$$\hat{\sigma}_x = \sigma_s \cdot \frac{1}{\sqrt{1 + \dfrac{2(s-r)}{r(s-2)(1+2\delta_{z_y})}}} \quad \text{where } \sigma_s = \sqrt{m_2}$$

This formula is, of course, subject to a systematic error that arises from the fact that we employ the square root of the best estimated 'most probable value' of the variance. Itmay be shown, however, that when r is large, the error is small.[14]

Consequently, we have the following theorem:

Theorem VIII. For a normal parent population, the best approximated 'most probable value' of the standard deviation of the parent population is equal to

$$\frac{\sigma_s}{\sqrt{1 + \dfrac{2(s-2r)}{r(s-2)(1+2\delta_{z_y})}}}$$

where σ_s is the standard deviation of an observed sample from the parent population and σ_{z_y} is a function of r and s as expressed in (37).

It is interesting to note from (39) that when $s = 2r$, $\hat{\sigma}_x = \sigma_s$ provided $\delta_{z_y} \neq -\frac{1}{2}$. However, from (37), δ_{z_y} cannot be equal to $-\frac{1}{2}$ in the case of $s = 2r$, where s and r are both positive integers. Consequently, we may state this fact in another theorem:

Theorem IX. If a sample is composed of exactly half of the variates of a normal parent population, then the best approximated 'most probable value' of the standard deviation of that parent population is equal to the standard deviation of an observed sample from it.

For an infinite parent population, (39) yields on reduction

$$\hat{\sigma}_x = \sigma_s \sqrt{\frac{r}{r+2}} \quad \text{for } \sigma_{z_y} = 0 \text{ when } s \to \infty. \tag{40}$$

---

[14] Professor H. C. Carver has worked out a relation between the most probable value of x^2 and that of x by assuming that the latter is distributed according to a Type III distribution. With his permission, I state the result as follows:

$$\text{M. P. V. } x^p = (\text{M. P. V. } x)^p \left(\frac{\lambda^2 - p}{\lambda^2 - 1}\right)^p$$

where $\lambda = \dfrac{M_x}{\sigma_x}$ and M_x = the distance of the mean from the origin.

Theorem X. For an infinite normal parent population, the best approximated 'most probable value' of the standard deviation of the parent population is equal to the standard deviation of an observed sample multiplied by $\sqrt{\bar{r}+2}$.

Section III. Distribution of the Hypothetical Means of the Parent Population

In the preceding two sections, we have obtained the best approximated 'most probable value' of the mean and the best approximated 'most probable value' of the standard deviation of a parent population assumed to be normal. We are now in the position to characterize the distribution of these hypothetical means by assuming that the best approximated 'most probable value' of the mean of the parent population be its mean and the best approximated 'most probable value' of the standard deviation of the parent population be its standard deviation. Such a characterization is subject to its own probable error.

Due to the fact that our parent population is normal by assumption, formulae (4), which we are to use this time, have to be modified by the proper substitution of the recursion relation of the moments of a normal distribution [See (21)]. After such modifications, they assume the following forms:

$$(41)\qquad \begin{cases} M_{z_x} = M_x \\ \bar{\mu}_{2:z_x} = \dfrac{s}{r^2} P_2 \bar{\mu}_{2:x} \\ \bar{\mu}_{3:z_x} = 0 \\ \bar{\mu}_{4:z_x} = \dfrac{3s}{r^4} (P_4 + P_2^2 s)\bar{\mu}_{2:x}^2 \\ \bar{\mu}_{5:z_x} = 0 \\ \bar{\mu}_{6:z_x} = \dfrac{15s}{r^6} (P_6 + 3P_4P_2 s + P_2^3 s^2)\bar{\mu}_{2:x}^3 \end{cases}$$

In accordance with Theorems II and III, we therefore have for the distribution of the means of the parent population the following:

$$(42)\qquad \begin{cases} M_{M_x} = m_1 \\ \bar{\mu}_{2:M_x} = \bar{\mu}_{2:z_x} = \dfrac{s}{r^2} P_2 \bar{\mu}_{2:x} = \dfrac{s}{r^2} P_2 \hat{\mu}_{2:x} \\ \bar{\mu}_{3:M_x} = -\bar{\mu}_{3:z_x} = 0 \\ \bar{\mu}_{4:M_x} = \bar{\mu}_{4:z_x} = \dfrac{3s}{r^4} (P_4 + P_2^2 s)\bar{\mu}_{2:x}^2 = \dfrac{3s}{r^4} (P_4 + P_2^2 s)\hat{\mu}_{2:x}^2 \\ \bar{\mu}_{5:M_x} = -\bar{\mu}_{5:z_x} = 0 \\ \bar{\mu}_{6:M_x} = \bar{\mu}_{6:z_x} = \dfrac{15s}{r^6} (P_6 + 3P_4P_2 s + P_2^3 s^2)\hat{\mu}_{2:x}^3 \\ \qquad = \dfrac{15s}{r^6} (P_6 + 3P_4P_2 s + P_2^3 s^2)\mu_{2:x}^3 \end{cases}$$

Consequently

$$(43)\qquad \begin{cases} M_{M_x} = m_1 \\ \sigma_{M_x} = \sqrt{\dfrac{s-r}{r(s-1)}}\,\hat{\sigma}_x \\ \alpha_{3:M_x} = 0 \\ \alpha_{4:M_x} - 3 = -\dfrac{6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)} \end{cases}$$

For an infinite parent population, i.e., $s \to \infty$, we have

$$(44)\qquad \begin{cases} M_{M_x} = m_1 \\ \sigma_{M_x} = \dfrac{1}{\sqrt{r}}\,\hat{\sigma}_x = \dfrac{1}{\sqrt{r}}\sqrt{\dfrac{r}{r+2}}\,\sigma_s = \dfrac{\sigma_s}{\sqrt{r+2}}, \text{ [from (40)]} \\ \alpha_{3:M_x} = 0 \\ \alpha_{4:M_x} - 3 = 0 \end{cases}$$

Now if we can find the equation of the curve associated with the distribution of the means of the parent population, we shall be able to ascertain the probability that a mean lies within certain limits after a sample from the parent population has once been observed.

Let us illustrate this by again referring to the same problem of the heights of 1000 freshman students as recorded in Table I. Considering this as our parent population which is almost normal with $s = 1000$, we take every tenth individual height from the original list in which the 1000 heights are tabulated. Thus we obtain a sample with $r = 100$. The frequency distribution of these 100 individual heights is shown in Table IV.

TABLE IV

Sample of 100 Heights Selected from the Parent Population of 1000 from Table I

Class	Frequency
62.5–64.4	9
64.5–66.4	16
66.5–68.4	31
68.5–70.4	29
70.5–72.4	13
72.5–74.4	2

We compute the mean, the standard deviation, the skewness, and the fourth moment about the mean of this sample:

$$\begin{aligned} m_1 &= 67.99 \\ m_2 &= 5.415,2 \qquad \sigma_s = 2.327,058 \\ m_3 &= -1.549,872 \qquad \alpha_{3:s} = -1.229,91 \\ m_4 &= 71.615,158 \qquad \alpha_{4:s} = 2.442,17 \end{aligned}$$

From Theorem VII,

$$\hat{M}_x = 67.99$$

From (37) and (39), we obtain

$$\begin{aligned} \delta_{xy} &= -.099,833 \\ \hat{\mu}_{2:x} &= 5.328,067 \end{aligned}$$

Substituting into (42), we have

$$\begin{aligned} M_{M_x} &= 67.99 \\ \bar{\mu}_{2:M_x} &= .048,000,603,6 \qquad \sigma_{M_x} = .219,09 \\ \bar{\mu}_{3:M_x} &= 0 \qquad \alpha_{3:M_x} = 0 \\ \bar{\mu}_{4:M_x} &= .006,898,429 \qquad \alpha_{4:M_x} = 2.994,03 \\ \bar{\mu}_{5:M_x} &= 0 \qquad \alpha_{5:M_x} = 0 \\ \bar{\mu}_{6:M_x} &= .001,649,027 \qquad \alpha_{6:M_x} = 14.910,37 \end{aligned}$$

The coefficients of Charlier's Type A Function (3) are as follows:

$$\frac{c_3}{3!} = 0$$

$$\frac{c_4}{4!} = -.000,250$$

$$\frac{c_5}{5!} = 0$$

$$\frac{c_6}{6!} = .000,000,1$$

From the values we are justified in assuming that M_x is normally distributed. We may now ask ourselves concerning the probability that the mean of the parent population, M_x, from which this sample is selected, exceeds 68.5 inches.

$$t = \frac{M_x - M_{M_x}}{\sigma_{M_x}} = \frac{68.5 - 67.99}{.21909} = 2.3278$$

$$P = \int_{2.3278}^{\infty} \phi(t)\, dt = .009962$$

Let us now come back to investigation of the general case for the distribution of the hypothetical means of the parent population. Because there is no definite relation between the values of r and s, except $r \leqq s$, and because, by assumption, our parent population is normal, δ_{z_x} is a function of r and s (22); that is

$$\delta_{z_x} = \frac{2s(r-1)(s-r-1)}{s(r-1)(s-r-1) - r(s-r)(s-2)(s-3)}$$

Consequently, it is necessary for us to investigate for different values of δ_{z_x} with respect to various combinations of r and s before we can tell which Type of Pearson's Curves will best fit the distribution of the means of the parent population. Hence, Table V:

TABLE V

Relation of the Values of δ_{z_x} with Various Combinations of r and s

$s = 10r$	$r \geqq 100$,	$\delta_{z_x} \geqq -.0020$
	$r \geqq 50$,	$\delta_{z_x} \geqq -.0040$
	$r \geqq 10$,	$\delta_{z_x} \geqq -.0189$
$s = 5r$	$r \geqq 100$,	$\delta_{z_x} \geqq -.0040$
	$r \geqq 50$,	$\delta_{z_x} \geqq -.0080$
	$r \geqq 10$,	$\delta_{z_x} \geqq -.0397$
$s = 2r$	$r \geqq 100$,	$\delta_{z_x} \geqq -.0101$
	$r \geqq 50$,	$\delta_{z_x} \geqq -.0204$
	$r \geqq 10$,	$\delta_{z_x} \geqq -.1118$

$s = r + 1$,	$r =$ any finite value,	$\delta_{z_x} = 0$
$s =$ any finite value,	$r = 1$	$\delta_{z_x} = 0$
$s \to \infty$,	$r =$ any finite value,	$\delta_{z_x} = 0$.

From the above table we observe:

1) For an infinite normal parent population, the frequency distribution of the hypothetical means of the parent population is normal, because both $\alpha_{3:M_x}$ and δ_{z_x} are equal to 0 (See Part I, Section III).

2) For any finite, normal parent population, if $r = 1$, the frequency distribution of the hypothetical means of the parent population is normal.

3) For any finite, normal parent population, if a sample $r = s - 1$ is chosen, the frequency distribution of the hypothetical means of the parent population is normal.

4) For any finite, normal parent population, if s is equal to $5r$ or more and at the same time r is at least equal to fifty, the normal curve is a fair approximation for the distribution of the hypothetical means of the parent population.

5) For the other cases in which $|\delta_{z_x}|$ is not negligibly small, we ought to make further investigation.

Now, to carry out further investigation for the cases where $|\delta_{z_x}|$ is not very small, we need only look back to formulae (43), from which we observe that: $\alpha_{4:M_x} - 3 < 0$ for $s \neq r + 1$, $r \neq 1$, or s does not approach infinity.

Because of the fact that $\alpha_{3:M_x} = 0$ and $\alpha_{4:M_x} < 3$ is the criterion for Type II,[15] we conclude that Type II will be the best fitting curve for the cases mentioned in 5) above. To obtain this Type II curve we proceed as follows:

Let the equation of the curve associated with the distribution of the hypothetical means of the parent population with which we are concerned be $y = P_{M_x}(t)$. Then

$$\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{b_0 + b_1 t + b_2 t^2} = \frac{a - t}{-b_2(t + R)(R - t)}$$

where

$$R = \frac{-b_1 \pm \sqrt{b_1^2 - 4b_0 b_2}}{2b_2}$$

By proper substitution with the formulae in Part I, Section III, we obatin

$$R = \frac{-\alpha_{3:M_x} \pm \sqrt{\alpha_{3:M_x}^2 - 4\delta_{z_x}(2 + \delta_{z_x})}}{2\delta_{z_x}} \tag{45}$$

$$= \pm \sqrt{\frac{-2}{\delta_{z_x}} - 1} \qquad \text{since } \alpha_{3:M_x} = 0 \text{ from (44)}$$

For the same reason $a = \frac{-\alpha_{3:M_x}}{2(1 + 2\delta_{z_x})} = 0$; therefore the differential equation may be rewritten as

$$\frac{1}{y}\frac{dy}{dt} = \frac{t}{b_2(R^2 - t^2)},$$

from which we obtain

$$y = y_0 (R^2 - t^2)^q \text{ where } q = -\frac{1}{2b_2} = -\frac{1 + 2\delta_{z_x}}{\delta_{z_x}}. \tag{46}$$

Imposing the condition that the total area under the curve be equal to unity, we set

$$1 = \int_{-R}^{R} y\,dt = y_0 \int_{-R}^{R} (R^2 - t^2)^q\,dt$$

[15] Elderton, W. P., *op. cit.*, Table VI, opposite p. 46.

Substituting $t = -R + 2R\mu$, we have

$$1 = y_0 \int_0^1 (2R)^{2q+1} \mu^q (1-\mu)^q \, du$$

$$= y_0 (2R)^{2q+1} \beta(q+1, q+1)$$

$$\therefore y_0 = \frac{1}{(2R)^{2q+1}} \cdot \frac{\Gamma(2q+2)}{\Gamma(q+1)\,\Gamma(q+1)}$$

hence

$$y = \frac{\Gamma(2q+2)}{(2R)^{2q+1}\,\Gamma(q+1)\,\Gamma(q+1)} (R^2 - t^2)^q$$

$$= \frac{1}{2^{2q+1}\sqrt{2q+3}} \cdot \frac{\Gamma(2q+2)}{\Gamma(q+1)\,\Gamma(q+1)} \left(1 - \frac{t^2}{2q+3}\right)^q, \tag{47}$$

where q may be expressed in terms of r and s by means of (46) and (22). Thus

$$q = -\frac{1}{\delta_{s_x}} - 2 = \frac{r(s-r)(s-2)(s-3) - 5s(r-1)(s-r-1)}{2s(r-1)(s-r-1)} \tag{48}$$

To sum up: In describing the distribution of the hypothetical means of a parent population from which our sample is chosen, we have the following theorems:

Theorem XI. The frequency distribution of the hypothetical means of an infinite, normal parent population is normal.

Theorem XII. The frequency distribution of the hypothetical means of a finite, normal parent population is normal if $r = s - 1$.

Theorem XIII. The frequency distribution of the hypothetical means of a finite, normal parent population is very nearly normal if s is equal to $5r$ or more and r is at least equal to fifty.

Theorem XIV. The frequency distribution of the hypothetical means of a finite, normal parent population is according to Type II for the cases in which $|\delta_{s_x}|$ is not negligibly small.

Section IV. Probable Error of the Mean

To measure the fluctuation of a sample mean from the true mean of the parent population, it is customary to use the term "probable error" to denote the expression:

$$E_M = 0.6745 \frac{\sigma_x}{\sqrt{r}} \tag{49}$$

where σ_x is the standard deviation of the parent population. As the true value of σ_x is not known, it is the common practice to substitute for it the value $\sqrt{\frac{r}{r-1}}\,\sigma_x'$, where σ_x' is the square root of the expected value of the sample second moment.

Therefore (49) is rewritten as

$$E_M = 0.6745 \frac{\sigma_x'}{\sqrt{r-1}} \tag{50}$$

Still, it should be noted, this expression is an approximation. Now from our theory of inverse sampling, as far as a normal parent population is assumed, we have obtained for the probable error of the mean

$$E_M = 0.6745 \frac{\sigma_s}{\sqrt{r+2}} \tag{51}$$

where σ_s is definitely the standard deviation of an observed sample. Although for large r, (50) and (51) do not differ much, yet (51) is obtained directly in terms of the standard deviation of an observed sample.

To illustrate, consider the same sample of the heights of 100 freshman students (See Table IV) as obtained from an infinite parent population. Since the mean is 67.99 and the standard deviation is 2.327058, the probable error of the mean is

$$E_M = 0.6745 \times \frac{2.327058}{\sqrt{102}} = .1554152;$$

that is, $M_x = 67.99 \pm .1554152$, which shows that the chances are even that the true mean of the parent population lies within the range 67.834,584,8 and 68.145,415,2.

Section V. Distribution of the Hypothetical Variances of the Parent Population

Recalling the fact we have stated in Part III, Section II, that the consideration of the distribution of the second moments of samples about the most probable value of the mean is equivalent to the consideration of a distribution of sample means drawn from a parent population $y_2, y_1, y_3, \cdots y_s$, where $y_i = (x_i - m_1)^2$ since in a normal parent population $\hat{M}_x = m_1$ [See (24)] we can write down in perfect analogy with (12) and (14)

$$\begin{aligned} \bar{\mu}_{n:p} &= (-1)^n \bar{\mu}_{n:z_y} \\ M_p &= m_2 \end{aligned} \tag{52}$$

Now

$$\bar{\mu}_{n:p} = \bar{\mu}_{n:M_y} = \bar{\mu}_{n:} \frac{\sum (x_i - m_1)^2}{N} = \bar{\mu}_{n:\bar{\mu}_{2::}}$$

since we have assumed the mean of the parent population to be its most probable value, i.e., m_1. Hence by virtue of (52) and (34), the frequency distribution of

the hypothetical variances of the parent population, which is assumed to be normal, is characterized by

$$(53)\qquad \begin{aligned} M_{\bar{\mu}_{2:x}} &= m_2 \\ \sigma_{\bar{\mu}_{2x}} = \sigma_{s_y} &= \sqrt{\frac{2(s-r)}{r(s-1)}}\,\sigma_x^2 = \sqrt{\frac{2(s-r)}{r(s-1)}}\,\hat{\sigma}_x^2 \end{aligned}$$

since we assume the most probable value of the variance of the parent population to be its variance.

$$\alpha_{3:\bar{\mu}_{2:x}} = -\alpha_{3:s_y} = -\frac{2(s-2r)}{s-2}\sqrt{\frac{2(s-1)}{r(s-r)}}$$

$$\alpha_{4:\bar{\mu}_{2:x}} - 3 = \alpha_{4:s_y} - 3 = \frac{12(s-1)(s^2+s-6rs+6r^2) - 6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)}$$

For an infinite parent population, i.e., $s \to \infty$, we have

$$(54)\qquad \begin{cases} M_{\bar{\mu}_{2:x}} = m_2 \\ \sigma_{\bar{\mu}_{2:x}} = \sqrt{\dfrac{2}{r}}\,\hat{\sigma}_x^2 \\ \alpha_{3.\bar{\mu}_{2}\,x} = -2\sqrt{\dfrac{2}{r}} \\ \alpha_{4:\bar{\mu}_{2:x}} - 3 = \dfrac{12}{r} \end{cases}$$

Now if we can find the equation of the curve associated with the distribution of the hypothetical variances of the parent population, we shall be able to ascertain the probability that a variance lies between certain specified limits after a sample is drawn from the parent population.

For illustration, we will use the same sample of the heights of 100 freshman students (See Table IV) as selected from a parent population of 1000.

We have $s = 1000, \quad r = 100$

$m_1 = 67.99$

$m_2 = 5.4152$, or $\sigma_s = 2.327058$

From (37) we compute

$$\delta_{s_y} = -.0098$$

As $|\delta_{s_y}|$ is negligibly small, we may be justified in considering $\bar{\mu}_{2:x}$ to be distributed according to Type III (Part I, Section III).

It follows from (39) that

$$\hat{\mu}_{2:x} = 5.32$$

We compute the moments of the distribution of the hypothetical variances in accordance with (53). Thus

$$M_{\bar{\mu}_{2:x}} = 5.4152$$
$$\sigma_{\bar{\mu}_{2:x}} = .556$$
$$\alpha_{3.\bar{\mu}_{2:x}} = .239,946$$
$$\alpha_{4:\bar{\mu}_{2:x}} = 3.055,75$$

If we now wish to ascertain the probability that the variance of the parent population lies between $\bar{\mu}_{2:x} = a = 5.5$ and $\bar{\mu}_{2:x} = b = 6.5$, we first convert a, b into standard units such that $t_a = .1525$ and $t_b = 1.9511$ and then evaluate the following integral:[16]

$$P = \frac{\left(\frac{2}{\alpha_3}\right)^{\frac{4}{\alpha_3^2}}}{\Gamma\left(\frac{4}{\alpha_3^2}\right)} e^{\frac{-4}{\alpha_3^2}} \int_{.1525}^{1.9511} \left(\frac{2}{\alpha_3} + t\right)^{\frac{4}{\alpha_3^2}-1} e^{-\frac{2}{\alpha_3}t} dt$$

But this step is now not necessary since we have access to Tables of Pearson's Type III Function.[17] Hence we find from this table our desired probability.

$$P = .39146$$

In the above numerical example, we are justified in using Type III because $|\delta_{z_y}|$ is negligibly small. But for the general case, however, we ought to make further investigation concerning the values of δ_{z_y}.

TABLE VI

Relation of Values of δ_{z_y} with Various Combinations of r and s

$s = 10r$	$r \geqq 100$	$\delta_{z_y} \geqq -.0098$
	$r \geqq 50$	$\delta_{z_y} \geqq -.0194$
	$r \geqq 10$	$\delta_{z_y} \geqq -.0859$
$s = 5r$	$r \geqq 100$	$\delta_{z_y} \geqq -.0200$
	$r \geqq 50$	$\delta_{z_y} \geqq -.0400$
	$r \geqq 10$	$\delta_{z_y} \geqq -.1983$
$s = 2r$	$r \geqq 100$	$\delta_{z_y} \geqq -.0518$
	$r \geqq 50$	$\delta_{z_y} \geqq -.1073$
	$r \geqq 10$	$\delta_{z_y} \geqq -.7642$

$s \to \infty$, r = any finite value, $\delta_{z_y} = 0$.

[16] Elderton, P. E., *op. cit.*, p. 90.

[17] Salvosa, L. R., Tables of Pearson's Type III Functions, *Annals of Mathematical Statistics* Vol. I, No. II.

Recalling that δ_{z_y} is a function of r and s such that

$$\delta_{z_y} = 2 - \frac{(s-3)\{4(s-2r)^2(s-1) + 2r(s-2)^2(s-r)\}}{(s-2)\{2(s-1)(s^2+s-6rs+6r^2) + r(s-r)(s-2)(s-3) - s(r-1)(s-r-1)\}}$$

we construct Table VI of δ_{z_y} for different combinations of s and r.

From Table VI we observe the following facts.

1) For an infinite, normal parent population, the distribution of the hypothetical variances of the parent population is according to Type III.

2) For a finite, normal parent population, if s is at least equal to $5r$ and r at least fifty, the distribution of the hypothetical variances of the parent population is very nearly according to Type III.

3) For the other cases in which δ_{z_y} is not small but negative in sign, the distribution of the hypothetical variances of the parent population needs further investigation.

From Part I, Section III, $k = \dfrac{\alpha_3^2}{4\delta(2+\delta)}$; and since we know that δ is always greater than -2, therefore whether k is positive or negative depends upon whether δ is positive or negative.

Now from Table VI we observe that δ_{z_y} seems to be always negative; hence k is negative. In accordance with the criterion for fitting curves, the frequency distribution of the variances of a normal parent population in such cases is according to Type I, which takes the form:[18]

$$y = \frac{\Gamma_{(m_1+m_2+2)}}{\Gamma_{(m_1+1)}\Gamma_{(m_2+1)}} \cdot \frac{1}{(R_1-R_2)^{m_1+m_2+1}} (t-R_2)^{m_1}(R_1-t)^{m_2} \tag{55}$$

where

$$m_1 = \frac{a-R_2}{b_2(R_2-R_1)}, \qquad m_2 = \frac{a-R_1}{b_2(R_1-R_2)}$$

R_1, R_2 are the positive and negative roots, respectively, of the equation $b_0 + b_1 t + b_2 t^2 = 0$ and can be expressed in terms of the first four moments:

$$R_1, R_2 = \frac{-\alpha_3 \pm \sqrt{\alpha_3^2 - 4\delta(2+\delta)}}{2\delta}$$

We may sum up the foregoing in the following theorems:

Theorem XV. The frequency distribution of the hypothetical variances of an infinite, normal parent population is according to Type III.

Theorem XVI. The frequency distribution of the hypothetical variances of a finite, normal parent population approximates to Type III Curve if r and s are of such combinations that $|\delta_{z_y}|$ turns out to be negligibly small.

Theorem XVII. The frequency distribution of the hypothetical variances of

[18] Elderton, W. P., *op. cit.*, p. 54.

a finite, normal parent population is according to Type I in case that δ_{z_y} is not very nearly equal to zero and is negative.

Part IV. Inverse Sampling Associated with a Parent Population Distributed According to Pearson's Type III Function

Instead of a normal parent population as we have assumed throughout our discussion in Part III, we shall assume in this part a parent population which is distributed according to Type III. Therefore, besides the distribution of the hypothetical means and that of the hypothetical variances of the parent population, the distribution of the hypothetical third moments will also be considered. We shall carry out our discussion in practically the same way as we have done in Part III.

Section I. Most Probable Value of the Mean of the Parent Population

We have already obtained a general expression for the most probable value of the mean of the parent population:

$$\hat{M}_x = m_1 - \frac{s - 2r}{r(s-2)} \frac{\sigma_x \alpha_{3:x}}{2(1 + 2\delta_{z_x})}$$

where as before

$$\delta_{z_x} = \frac{2\alpha_{4:z_x} - 3\alpha^2_{3:z_x} - 6}{\alpha_{4:z_x} + 3}$$

But we are now concerned with a parent population which is distributed according to Type III.

Since the recursion relation of the moments of Type III distribution is of the form

$$\alpha_{n+1} = n\left(\alpha_{n-1} + \frac{\alpha_3}{2}\alpha_n\right) \tag{56}$$

$\alpha_4 = 3(1 + \gamma)$ where $\gamma = \frac{\alpha_3^2}{2}$

$\alpha_5 = 2\alpha_3(5 + 3\gamma)$

$\alpha_6 = 5(3 + 13\gamma + 6\gamma^2)$

$\alpha_7 = 3\alpha_3(35 + 77\gamma + 30\gamma^2)$

$\alpha_8 = 7(15 + 170\gamma + 261\gamma^2 + 90\gamma^3)$

$\alpha_9 = 4\alpha_3(315 + 1652\gamma + 2007\gamma^2 + 630\gamma^3)$

$\alpha_{10} = 9(105 + 2450\gamma + 8435\gamma^2 + 8658\gamma^3 + 2520\gamma^4)$

$\alpha_{11} = 5\alpha_3(3456 + 35266\gamma + 91971\gamma^2 + 82962\gamma^3 + 22680\gamma^4)$

$\alpha_{12} = 11(945 + 39375\gamma + 252245\gamma^2 + 537777\gamma^3 + 437490\gamma^4 + 113400\gamma^5)$

etc.

it follows from (5) that for a Type III distribution of the parent population

$$(57)\quad \begin{cases} M_{s_x} = M_x \\ \sigma_{s_x} = \sqrt{\dfrac{s-r}{r(s-1)}}\,\sigma_x \\ \alpha_{3:s_x} = \dfrac{s-2r}{s-2}\sqrt{\dfrac{s-1}{r(s-r)}}\,\alpha_{3:x} \\ \alpha_{4:s_x} - 3 = \dfrac{(s-1)(s^2+s-6rs+6r^2)}{r(s-r)(s-2)(s-3)}\cdot\dfrac{\alpha_{3:x}^2}{2} - \dfrac{6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)} \end{cases}$$

Therefore for the most probable value of the mean of the parent population, we have the same form as (20):

$$\hat{M}_x = m_1 - \frac{s-2r}{r(s-2)}\,\frac{\sigma_x \alpha_{3:x}}{2(1+2\delta_{s_x})};$$

except now instead of (17)

$$(58)\quad \delta_{s_x} = 2 - \frac{3\alpha_{3:s_x}^2 + 12}{\alpha_{4:s_x}+3}$$

$$= 2 - \frac{(s-3)\{2(s-1)(s-2r)^2\alpha_{3:x}^2 + 8r(s-2)^2(s-r)\}}{(s-2)\{(s-1)(s^2+s-6rs+6r^2)\alpha_{3:x}^2 + 4r(s-2)(s-3)(s-r) - 4s(r-1)(s-r-1)\}}$$

We observe that if $\alpha_{3:x} = 0$, this comes back to the case of normal parent population which we have already treated in Part III.

But if $s \to \infty$ while $\alpha_{3:x}$ is finite, then $\delta_{s_x} = 0$. Therefore, for the limiting case, i.e., when the parent population is infinite, we have

$$(59)\quad \hat{M}_x = m_1 - \frac{1}{2r}\,\sigma_x\,\alpha_{3:x}$$

Since σ_x and $\alpha_{3:x}$ are not known, we impose the condition that they assume their best approximated 'most probable values' respectively. Hence, we rewrite (20), (59) in the following forms:

$$(60)\quad \hat{M}_x = m_1 - \frac{s-2r}{r(s-2)}\,\frac{\hat{\sigma}_x\,\hat{\alpha}_{3:x}}{2(1+2\hat{\delta}_{s_x})}$$

where now

$$(60b)\quad \hat{\delta}_{s_x} = 2 - \frac{(s-3)\{2(s-1)(s-2r)^2\hat{\alpha}_{3:x}^2 + 8r(s-2)^2(s-r)\}}{(s-2)\{(s-1)(s^2+s-6rs+6r^2)\hat{\alpha}_{3:x}^2 + 4r(s-2)(s-3)(s-r) - 4s(r-1)(s-r-1)\}}$$

and for the infinite case

$$\hat{M}_x = m_1 - \frac{1}{2r}\,\hat{\sigma}_x\,\hat{\alpha}_{3:x} \tag{61}$$

So we state our theorem:

Theorem XVIII. For a parent population which is distributed according to Type III, the best approximated 'most probable value' of the mean is the mean of an observed sample from it minus a correction factor which is a function of r, s, $\hat{\sigma}_x$, and $\hat{\alpha}_{3:x}$.

It is also interesting to note that when $s = 2r$ and $\hat{\delta}_{x_r} \neq -\frac{1}{2}$, then $\hat{M} = m_1$.

Section II. Most Probable Value of the Standard Deviation of the Parent Population

We consider, as we have done in Part III, Section II, ${}_sC_r$ possible samples, each consisting of r variates chosen from a parent population s. The second moment of each sample computed about the most probable value of the mean of the parent population may be written as

$$z_1 = \frac{1}{r}\{(x_1 - \hat{M}_x)^2 + (x_2 - \hat{M}_x)^2 + \cdots + (x_r - \hat{M}_x)^2\}$$

$$z_2 = \frac{1}{r}\{(x_2 - \hat{M}_x)^2 + (x_3 - \hat{M}_x)^2 + \cdots + (x_{r+1} - \hat{M}_x)^2\}$$

. . . . . .

$$z_{\binom{s}{r}} = \frac{1}{r}\{(x_{s-r+1} - \hat{M}_x)^2 + (x_{s-r+2} - \hat{M}_x)^2 + \cdots + (x_s - \hat{M}_x)^2\}$$

If we write $(x_i - \hat{M}_x)^2 = y_i$, the above may be considered as a distribution of sample means drawn from a parent population $y_1, y_2, y_3, \cdots y_s$. Therefore, as (27),

$$\mu_{n:y} = \bar{\mu}_{2n:x} + \binom{2n}{1}\bar{\mu}_{2n-1:x}(M_x - \hat{M}_x) + \binom{2n}{2}\bar{\mu}_{2n-2:x}(M_x - \hat{M}_x)^2$$

$$+ \binom{2n}{3}\bar{\mu}_{2n-3:x}(M_x - \hat{M}_x)^3 + \cdots + (M_x - \hat{M}_x)^{2n}$$

When we impose the condition that the most probable value of the mean of the parent population be its mean, then the above yields

$$\mu_{n:y} = \bar{\mu}_{2n:x}$$

$$\mu_{1:y} = M_y = \bar{\mu}_{2:x}$$

Consequently

$$\bar{\mu}_{k:y} = \frac{\sum (y - M_y)^k}{N} = \mu_{k:y} - \binom{k}{1}\mu_{k-1:y} M_y + \binom{k}{2}\mu_{k-2:y} M_y^2 - \cdots + (-1)^k M_y^k$$

$$= \bar{\mu}_{2k:x} - \binom{k}{1}\bar{\mu}_{2k-2:x}\bar{\mu}_{2:x} - \binom{k}{2}\bar{\mu}_{2k-4:x}\bar{\mu}_{2:x}^2 - \cdots + (-1)^k \bar{\mu}_{2:x}^k.$$

Now from the fact that we assume a Type III distribution for our parent population, therefore we have

$$(62)\quad \begin{cases} \bar{\mu}_{2:y} = \bar{\mu}_{4:x} - \bar{\mu}_{2:x}^2 = (3\gamma + 2)\sigma_x^4 \\ \bar{\mu}_{3:y} = \bar{\mu}_{6:x} - 3\bar{\mu}_{4:x}\bar{\mu}_{2:x} + 2\bar{\mu}_{2:x}^3 = (30\gamma^2 + 56\gamma + 8)\sigma_x^6 \\ \bar{\mu}_{4:y} = \bar{\mu}_{8:x} - 4\bar{\mu}_{6:x}\bar{\mu}_{2:x} + 6\bar{\mu}_{4:x}\bar{\mu}_{2:x}^2 - 3\bar{\mu}_{2:x}^4 \\ \qquad = (630\gamma^3 + 1707\gamma^2 + 948\gamma + 60)\sigma_x^8 \\ \qquad\qquad \text{etc.} \end{cases}$$

Substituting (62) into (26), we have

$$(63)\quad \begin{cases} M_{z_y} = \sigma_x^2 \\ \sigma_{z_y} = \sigma_x^2 \sqrt{\dfrac{s-r}{r(s-1)}(3\gamma + 2)} \\ \alpha_{3:z_y} = \dfrac{s-2r}{s-2}\sqrt{\dfrac{s-1}{r(s-r)}} \cdot \dfrac{30\gamma^2 + 56\gamma + 8}{(3\gamma + 2)^{3/2}} \\ \alpha_{4:z_y} - 3 = \dfrac{(s-1)(s^2 + s - 6rs + 6r^2)}{r(s-r)(s-2)(s-3)} \cdot \dfrac{630\gamma^3 + 1680\gamma^2 + 912\gamma + 48}{(3\gamma + 2)^2} \\ \qquad - \dfrac{6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)}. \end{cases}$$

For an infinite parent population, the above yields by allowing $s \to \infty$

$$(64)\quad \begin{cases} M_{z_y} = \sigma_x^2 \\ \sigma_{z_y} = \sigma_x^2 \sqrt{\dfrac{3\gamma + 2}{r}} \\ \alpha_{3:z_y} = \dfrac{1}{\sqrt{r}} \dfrac{30\gamma^2 + 56\gamma + 8}{(3\gamma + 2)^{3/2}} \\ \alpha_{4:z_y} - 3 = \dfrac{1}{r} \dfrac{630\gamma^3 + 1680\gamma^2 + 912\gamma + 48}{(3\gamma + 2)^2} \end{cases}$$

In accordance with (38), we write

$$(65)\qquad \frac{z_y - \sigma_x^2}{\sigma_x^2 \sqrt{\frac{s-r}{r(s-1)}(3\gamma+2)}} = \frac{-\frac{s-2r}{s-2}\sqrt{\frac{s-1}{r(s-r)}}\frac{30\gamma^2+56\gamma+8}{(3\gamma+2)^{3/2}}}{2(1+2\delta_{z_y})}$$

It follows from Theorem IV that for the mode of the standard deviation of the parent population, we have

$$(66)\qquad \frac{\frac{\sum (x-\hat{M}_x)^2}{r} - \hat{\sigma}_x^2}{\hat{\sigma}_x^2 \sqrt{\frac{s-r}{r(s-1)}(3\gamma+2)}} = \frac{\frac{s-2r}{s-2}\sqrt{\frac{s-1}{r(s-r)}}\frac{30\gamma^2+56\gamma+8}{(3\gamma+2)^{\frac{3}{2}}}}{2(1+2\delta_{z_y})}$$

where

$$(67)\qquad \delta_{z_y} = \frac{2\alpha_{4:z_y} - 3\alpha_{3:z_y}^2 - 6}{\alpha_{4:z_y} + 3}$$

which is a function of r, s, and $\alpha_{3:x}$.

Assuming the best approximated 'most probable value' of $\alpha_{3:x}$ for $\alpha_{3:x}$ and remembering that

$$\hat{M}_x = m_1 - \frac{s-2r}{r(s-2)}\frac{\hat{\sigma}_x \hat{\alpha}_{3:x}}{2(1+2\hat{\delta}_{z_x})},$$

we write (66) in the form of

$$m_2 + g^2 \frac{\hat{\sigma}_x^2 \alpha_{3:x}^2}{(1+2\hat{\delta}_{z_x})^2} - \hat{\sigma}_x^2 = g\frac{30\hat{\gamma}^2+56\hat{\gamma}+8}{(3\hat{\gamma}+2)(1+2\hat{\delta}_{z_y})}\hat{\sigma}_x^2$$

$$(68)\qquad \hat{\sigma}_x^2 = \frac{m_2}{1 + g\frac{30\hat{\gamma}^2+56\hat{\gamma}+8}{(3\hat{\gamma}+2)(1+2\hat{\delta}_{z_y})} - \frac{2g^2\hat{\gamma}}{(1+2\hat{\delta}_{z_x})^2}}$$

where

$$g = \frac{s-2r}{2r(s-2)},$$

$\hat{\delta}_{z_x}$ = (60.b) where $\alpha_{3:x}$ is replaced by $\hat{\alpha}_{3:x}$

$\hat{\delta}_{z_y}$ = (67) where $\alpha_{3:x}$ is replaced by $\hat{\alpha}_{3:x}$

$$\gamma = \frac{\hat{\alpha}_{3:x}^2}{2}$$

We rewrite (68) in the abridged form:

$$\hat{\sigma}_x^2 = \frac{m_2}{\phi^2(\hat{\alpha}_{3:x}, r, s)} \tag{69}$$

or

$$\hat{\sigma}_x = \frac{\sigma_s}{\phi(\hat{\alpha}_{3:x}, r, s)}$$

where

$$\phi(\hat{\alpha}_{3:x}, r, s) = \sqrt{1 + g\frac{30\hat{\gamma}^2 + 56\hat{\gamma} + 8}{(3\hat{\gamma} + 2)(1 + 2\hat{\delta}_{z_y})} - \frac{2g^2\hat{\gamma}}{(1 + 2\hat{\delta}_{z_x})^2}}$$

and state our theorem:

Theorem XIX. The best approximated 'most probable value' of the standard deviation of a parent population which is assumed to be distributed according to Type III is equal to the standard deviation of an observed sample of it, multiplied by $\frac{1}{\phi(\hat{\alpha}_{3:x}, r, s)}$.

For an infinite parent population, $g = \frac{1}{2r}$, $\hat{\delta}_{z_x} = 0$ and

$$\hat{\delta}_{z_y} = \frac{2(630\hat{\gamma}^3 + 1680\hat{\gamma}^2 + 912\hat{\gamma} + 48)(3\hat{\gamma} + 2) - 3(30\hat{\gamma}^2 + 56\hat{\gamma} + 8)^2}{(3\hat{\gamma} + 2)[(630\hat{\gamma}^3 + 1680\hat{\gamma}^2 + 912\hat{\gamma} + 48) - 6r(3\hat{\gamma} + 2)^2]}$$

Theorem XX. The best approximated 'most probable value' of the standard deviation of an infinite parent population which is assumed to be distributed according to Type III is equal to the standard deviation of an observed sample of it, multiplied by $\frac{1}{\lim_{s\to\infty} \phi(\hat{\alpha}_{3:x}, r, s)}$.

Section III. Most Probable Value of the Skewness of the Parent Population

Let us again consider ${}_sC_r$ samples, each consisting of r variates chosen from a parent population s. The third moments of each sample computed about the most probable value of the mean of the parent population may be written as

$$\begin{aligned}
Z_1 &= \frac{1}{r}\{(x_1 - \hat{M}_x)^3 + (x_2 - \hat{M}_x)^3 + \cdots + (x_r - \hat{M}_x)^3\} \\
Z_2 &= \frac{1}{r}\{(x_2 - \hat{M}_x)^3 + (x_3 - \hat{M}_x)^3 + \cdots + (x_{r+1} - \hat{M}_x)^3\} \qquad (70) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
Z_{\binom{s}{r}} &= \frac{1}{r}\{(x_{s-r+1} - \hat{M}_x)^3 + (x_{s-r+2} - \hat{M}_x)^3 + \cdots + (x_s - \hat{M}_x)^3\}
\end{aligned}$$

If we write $(x_i - \hat{M}_x)^3 = w_i$, the above may be considered as a distribution of sample means drawn from a parent population $w_1, w_2, w_3, \cdots w_s$. Consequently in accordance with (5), we have

$$(71)\qquad \begin{cases} M_{z_w} = M_w \\ \sigma_{z_w} = \sqrt{\dfrac{s-r}{r(s-1)}}\,\sigma_w \\ \alpha_{3:z_w} = \dfrac{s-2r}{s-2}\sqrt{\dfrac{s-1}{r(s-r)}}\,\alpha_{3:w} \\ \alpha_{4:z_w} - 3 = \dfrac{(s-1)(s^2+s-6rs+6r^2)}{r(s-r)(s-2)(s-3)}\{\alpha_{4:w}-3\} \\ \qquad\qquad - \dfrac{6s(r-1)(s-r-1)}{r(s-1)(s-2)(s-3)}. \end{cases}$$

Let us write the analogous form of (27):

$$(72)\qquad \begin{aligned} \mu_{n:w} &= \frac{1}{N}\sum w^n = \frac{1}{N}\sum (x - M_x + M_x - \hat{M}_x)^{3n} \\ &= \bar{\mu}_{3n:x} + \binom{3n}{1}\bar{\mu}_{3n-1:x}(M_x - \hat{M}_x) \\ &\quad + \binom{3n}{2}\bar{\mu}_{3n-2:x}(M_x - \hat{M}_x)^2 + \cdots + (M_x - \hat{M}_x)^{3n} \end{aligned}$$

Imposing the same condition as before that M_x assumes its most probable value (i.e., $M_x = \hat{M}_x$), then (72) becomes

$$(73)\qquad \begin{aligned} \mu_{n:w} &= \bar{\mu}_{3n:x} \\ \mu_{1:w} &= M_\omega = \bar{\mu}_{3:x} \end{aligned}$$

The kth moment of the distribution of w about its mean will then be

$$(74)\qquad \begin{aligned} \bar{\mu}_{k:w} &= \frac{\sum (w - M_w)^k}{N} = \mu_{k:w} - \binom{k}{1}\mu_{k-1:w}\,M_w \\ &\quad + \binom{k}{2}\mu_{k-2:w}\,M_w^2 - \cdots + (-1)^k M_w^k \\ &= \bar{\mu}_{3k:x} - \binom{k}{1}\bar{\mu}_{3k-3:x}\,\bar{\mu}_{3:x} + \binom{k}{2}\bar{\mu}_{3k-6:x}\,\bar{\mu}_{3:x}^2 \\ &\quad - \cdots + (-1)^k \bar{\mu}_{3:x}^k \end{aligned}$$

Since we assume a Type III distribution of the parent population, we have in accordance with the recursion relation (56)

$$
\begin{aligned}
M_w &= \bar{\mu}_{3:x} = \alpha_{3:x}\,\sigma_x^3 \\
\bar{\mu}_{2:w} &= \bar{\mu}_{6:x} - \bar{\mu}_{3:x}^2 = (15 + 63\gamma + 30\gamma^2)\,\sigma_x^6 \\
\bar{\mu}_{3:w} &= \bar{\mu}_{9:x} - 3\bar{\mu}_{6:x}\,\bar{\mu}_{3:x} + 2\bar{\mu}_{3:x}^3 \\
&= \alpha_{3:x}\,(1215 + 6417\gamma + 7938\gamma^2 + 2520\gamma^3)\,\sigma_x^9 \\
\bar{\mu}_{4:w} &= \bar{\mu}_{12:x} - 4\bar{\mu}_{9:x}\,\bar{\mu}_{3:x} + 6\bar{\mu}_{6:x}\,\bar{\mu}_{3:x}^2 - 3\bar{\mu}_{3:x}^4 \\
&= (10395 + 423225\gamma + 2722599\gamma^2 + 5851683\gamma^3 \\
&\quad + 4792230\gamma^4 + 1247400\gamma^5)\,\sigma_x^{12}
\end{aligned}
\tag{75}
$$

Substituting into (71), we have

$$
\begin{cases}
M_{z_w} = \bar{\mu}_{3:x} = \alpha_{3:x}\,\sigma_x^3 \\
\sigma_{z_w} = \sigma_x^3 \sqrt{\dfrac{s-r}{r(s-1)}\,(30\gamma^2 + 63\gamma + 15)} \\
\alpha_{3:z_w} = \dfrac{s-2r}{s-2}\sqrt{\dfrac{s-1}{r(s-r)}}\,\dfrac{\alpha_{3,y}\,(1215 + 6417\gamma + 7938\gamma^2 + 2520\gamma^3)}{(15 + 63\gamma + 30\gamma^2)^{3/2}} \\
\alpha_{4:z_w} - 3 = \dfrac{(s-1)\,(s^2 + s - 6rs + 6r^2)}{r(s-r)\,(s-2)\,(s-3)} \\
\qquad \cdot \dfrac{9720 + 417555\gamma + 2707992\gamma^2 + 5840343\gamma^3 + 4789530\gamma^4 + 1247400\gamma^5}{(15 + 63\gamma + 30\gamma^2)^2} \\
\qquad - \dfrac{6s(r-1)\,(s-r-1)}{r(s-r)\,(s-2)\,(s-3)}.
\end{cases}
\tag{76}
$$

Allowing $s \to \infty$, we have for an infinite parent population

$$
\begin{aligned}
M_{z_w} &= \bar{\mu}_{3:x} = \alpha_{3:x}\,\sigma_x^3 \\
\sigma_{z_w} &= \sigma_x^3 \sqrt{\frac{1}{r}\,(30\gamma^2 + 63\gamma + 15)} \\
\alpha_{3:z_w} &= \frac{1}{\sqrt{r}}\,\frac{\alpha_{3:x}\,(1215 + 6417\gamma + 7938\gamma^2 + 2520\gamma^3)}{(15 + 63\gamma + 30\gamma^2)^{3/2}} \\
\alpha_{4:z_w} - 3 &= \frac{1}{r} \cdot \frac{9720 + 417555\gamma + 2707992\gamma^2 + 5840343\gamma^3 + 4789530\gamma^4 + 1247400\gamma^5}{(15 + 63\gamma + 30\gamma^2)^2}
\end{aligned}
\tag{77}
$$

The best approximated 'most probable value' of $\bar{\mu}_{3:x}$ may now be written after the same fashion as in the preceding cases:

$$\hat{\mu}_{3:x} = \frac{\sum (x - \hat{M}_x)^3}{r} - \frac{\sigma_{z_w} \cdot \alpha_{3:z_w}}{2(1 + 2\delta_{z_w})} \tag{78}$$

where

$$\delta_{z_w} = \frac{2\alpha_{4:z_w} - 3\alpha^2_{3:z_w} - 6}{\alpha_{4:z_w} + 3}$$

Since

$$\frac{\sum (x - \hat{M}_x)^3}{r} = \frac{1}{r} \sum \left(x - m_1 + g \frac{\hat{\sigma}_x \hat{\alpha}_{3:x}}{1 + 2\hat{\delta}_{z_x}} \right)^3 \qquad \text{[from (60)]},$$

and since we assume the best approximated 'most probable values' of the standard deviation and the skewness for the standard deviation and the skewness of the parent population respectively, we obtain from (78)

$$\hat{\sigma}_x^3 \hat{\alpha}_{3:x} = m_3 + 3m_2 g \frac{\hat{\sigma}_x \hat{\alpha}_{3:x}}{1 + 2\hat{\delta}_{z_x}} + g^3 \frac{\hat{\sigma}_x^3 \hat{\alpha}^3_{3:x}}{(1 + 2\hat{\delta}_{z_x})^3}$$

$$- g \frac{\hat{\alpha}_{3:x} (1215 + 6417\hat{\gamma} + 7938\hat{\gamma}^2 + 2520\hat{\gamma}^3)}{(1 + 2\hat{\delta}_{z_w})(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)} \hat{\sigma}_x^3$$

The change of $\hat{\mu}_{3:x}$ to $\hat{\sigma}_x^3 \hat{\alpha}_{3:x}$ involves a systematic error although it is small.

Again by proper substitution of (69) we have

$$\frac{\sigma_3^3 \hat{\alpha}_{3:x}}{\phi^3(\hat{\alpha}_{3:x}, r, s)} = \sigma_s^3 \alpha_{3:s} + 3\sigma_s^3 g \frac{\hat{\alpha}_{3:x}}{\phi(\hat{\alpha}_{3:x}, r, s)(1 + 2\hat{\delta}_{z_x})}$$

$$+ g^3 \frac{\sigma_s^3 \hat{\alpha}^3_{3:x}}{\phi^3(\hat{\alpha}_{3:x}, r, s)(1 + 2\hat{\delta}_{z_x})^3}$$

$$- g \frac{\hat{\alpha}_{3:x} \sigma_s^3 (1215 + 6417\hat{\gamma} + 7938\hat{\gamma}^2 + 2520\hat{\gamma}^3)}{\phi^3(\hat{\alpha}_{3:x}, r, s) \cdot (15 + 63\hat{\gamma} + 30\hat{\gamma}^2)(1 + 2\hat{\delta}_{z_w})}.$$

Solving for $\alpha_{3:s}$, we have

$$\alpha_{3:s} = \frac{\hat{\alpha}_{3:x}}{\phi^3(\hat{\alpha}_{3:x}, r, s)} \left[1 - \frac{3g\phi^2(\hat{\alpha}_{3:x}, r, s)}{(1 + 2\hat{\delta}_{z_x})} - \frac{2\hat{\gamma} g^3}{(1 + 2\hat{\delta}_{z_x})^3} \right.$$

$$\left. + \frac{g(1215 + 6417\hat{\gamma} + 7938\hat{\gamma}^2 + 2520\hat{\gamma}^3)}{(1 + 2\hat{\delta}_{z_w})(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)} \right]. \tag{79}$$

Since the right member of (79) is a function of $\hat{\alpha}_{3:x}$, r, and s, therefore the most probable value of $\alpha_{3:x}$ may be approximated when we are given s, r, and the skewness of an observed sample. As it is an algebraic equation of high order in $\hat{\alpha}_{3:x}$ and is so much involved, even approximation presents practical

difficulty. However, if once $\hat{\alpha}_{3:x}$ is approximated, $\hat{\sigma}_x$ and $\hat{M}_x$ can be easily obtained from (60) and (68).

Theorem XXI. For the best approximated 'most probable value' of the skewness of a parent population which is assumed to be distributed according to Type III, we must approximate it from equation (79), in which the skewness of an observed sample is expressed as a function of s, r, and the best approximated 'most probable value' of the skewness of the parent population.

To construct a table for the best approximated 'most probable value' $\hat{\alpha}_{3:x}$ corresponding to $\alpha_{3:s}$ for particular values of r, s, we should first reverse the process by assigning different values of $\hat{\alpha}_{3:x}$ so as to obtain $\alpha_{3:s}$; then by the way of interpolation, we shall be able to obtain $\hat{\alpha}_{3:x}$ for a particular $\alpha_{3:s}$.

TABLE VII

Relation of the Sample Skewness and the Best Approximated 'Most Probable Value' of the Parent Population Whose Distribution is According to Type III

($s \to \infty$, $r = 100$)

$\alpha_{3:s}$	$\hat{\alpha}_{3:x}$
.1	.0784
.2	.1568
.3	.2373
.4	.3164
.5	.3969
.6	.4776
.7	.5589
.8	.6410
.9	.7239
1.0	.8072
1.1	.8905
1.2	.9737
1.3	1.0567
1.4	1.1392
1.5	1.2211
1.6	1.3022
1.7	1.3791
1.8	1.4578
1.9	1.5355
2.0	1.6122
2.1	1.6828
2.2	1.7609
2.3	1.8303
2.4	1.9024
2.5	1.9670
2.6	2.0371

For $s \to \infty$ and $r = 100$, we have computed the best approximated 'most probable value' of $\alpha_{3:x}$ corresponding to the values of $\alpha_{3:s}$ from .1 to 2.6 as shown in Table VII.

The computation for such a table is laborious because it involves the computation of $\hat{\delta}_{z_x}$, $\hat{\delta}_{z_y}$, and $\hat{\delta}_{z_\omega}$ which are in turn functions of $\hat{\alpha}_{3:z_x}$ and $\hat{\alpha}_{4:z_x}$, $\hat{\alpha}_{3:z_y}$ and $\hat{\alpha}_{4:z_y}$, and $\hat{\alpha}_{3:z_\omega}$ and $\hat{\alpha}_{4:z_\omega}$, respectively.

Section IV. Distribution of the Hypothetical Means of the Parent Population

Since we have obtained in the preceding sections expressions for the best approximated 'most probable values' of the mean, the standard deviation and the skewness of a parent population which is assumed to be distributed according to Type III, we are now in the position to characterize the distribution of the hypothetical means of the parent population with the assumption that the best approximated 'most probable values' of the mean, the standard deviation, and the skewness be the mean, the standard deviation, and the skewness of the parent population.

Basing upon the fundamental relations in (15), we write down the characteristics of the distribution of the hypothetical means of the parent population as follows:

$$(80)\quad \begin{cases} M_{M_x} = m_1 \\ \sigma_{M_x} = \sigma_{z_x} = \hat{\sigma}_x \sqrt{\dfrac{s-r}{r(s-1)}} = \dfrac{\sigma_s}{\phi(\hat{\alpha}_{3:x}, s, r)} \sqrt{\dfrac{s-1}{r(s-1)}} \\ \alpha_{3:M_x} = -\alpha_{3:z_x} = -\dfrac{s-2r}{s-2} \sqrt{\dfrac{s-1}{r(s-r)}}\, \hat{\alpha}_{3:x} \\ \alpha_{4:M_x} - 3 = \alpha_{4:z_x} - 3 \\ \qquad = \dfrac{(s-1)(s^2+s-6rs+6r^2)}{r(s-r)(s-2)(s-3)} \left[\dfrac{3\hat{\alpha}^2_{3:x}}{2}\right] - \dfrac{6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)}. \end{cases}$$

where $\phi(\hat{\alpha}_{3:x}, s, r)$ is given in (69).

For an infinite parent population by allowing $s \to \infty$, we obtain from the above:

$$(81)\quad \begin{cases} M_{M_x} = m_1 \\ \sigma_{M_x} = \dfrac{1}{\sqrt{r}} \dfrac{\sigma_s}{\phi(\hat{\alpha}_{3:x}, r)} \\ \alpha_{3:M_x} = \dfrac{1}{\sqrt{r}}\, \hat{\alpha}_{3:x} \\ \alpha_{4:M_x} - 3 = \dfrac{3}{2r}\, \hat{\alpha}^2_{3:x} \end{cases}$$

where $\phi(\hat{\alpha}_{3:x}, r) = \lim_{s \to \infty} \phi(\hat{\alpha}_{3:x}, s, r)$

Since we observe that the moments of the distribution of the hypothetical means are expressed in terms of $\hat{\alpha}_{3:x}$, it is therefore necessary for us to find the best approximated 'most probable value' of the skewness of a parent population before we attempt to obtain the frequency function associated with the distribution of these hypothetical means.

Numerical illustration. A sample of 100 weights of freshman students is observed and the frequency distribution is given in Table VIII.

TABLE VIII

Weights of 100 Freshman Students

(Original Measurements Correct to Nearest Pound)

Class Mark	Frequency
109.5	4
119.5	11
129.5	25
139.5	34
149.5	14
159.5	8
169.5	0
179.5	3
189.5	1
	100

The first four moments are computed

$$\begin{aligned} m_1 &= 138.3 \\ \sigma_s &= 14.6366 \\ \alpha_{3:s} &= .81099 \\ \alpha_{4:s} &= 4.47644 \end{aligned}$$

Now, assuming this sample is drawn from an infinite parent population which is assumed to be distributed according to Type III, we wish to find (a) the best approximated 'most probable values' of the mean, the standard deviation, and the skewness of the parent population, and (b) the probability that the mean of the parent population lies between $M_x = 135$ and $M_x = 140$.

By interpolation from Table VII, we obtain the best approximated 'most probable value' of the skewness of the parent population:

$$\hat{\alpha}_{3:x} = .6501$$

From (69) and (61) we obtain

$$\begin{aligned} \hat{\sigma}_x &\doteq 14.5452 \\ \hat{M}_x &= 138.25272, \qquad \phi(\hat{\alpha}_{3:x}, r) = 1.006279 \end{aligned}$$

From (81) we have

$$\begin{aligned} M_{M_x} &= 138.3 \\ \sigma_{M_x} &= 1.45452 \\ \alpha_{3:M_x} &= .06501 \\ \alpha_{4:M_x} &= 3.00633945 \end{aligned}$$

$\delta_{z_x} = 0$, the distribution of M_x is associated with Type III Function; hence for the probability that M_x lies between $M_x = 135$ and $M_x = 140$, we again refer to Tables of Pearson's Type III Function prepared by L. R. Salvosa,[19] and we obtain in this case

$$P = .8677592$$

Since the determination of the best fit of a frequency curve in general depends upon the values of α_3, α_4, and k, and since in the present case each of them is a function of s, r, and $\alpha_{3:x}$, we are therefore not able to tell the type of curve to be used until we know s, r, and $\hat{\alpha}_{3:x}$.

For the infinite case, however, as we have illustrated Type III Function may always be used because

$$\delta_{z_x} = \frac{2\alpha_{4:z_x} - 3\alpha_{3:z_x}^2 - 6}{\alpha_{4:z_x} + 3} = \frac{2\alpha_{4:M_x} - 3\alpha_{3:M_x}^2 - 6}{\alpha_{4:M_x} + 3} = 0$$

holds for all values of $\hat{\alpha}_{3:x}$ and r. We therefore conclude that the hypothetical means of an infinite parent population which is itself distributed according to Type III is distributed according to Type III. Hence

Theorem XXII. The hypothetical means of an infinite parent population is distributed according to Type III if the parent population is assumed to be distributed according to Type III.

Section V. Distribution of the Hypothetical Variances of the Parent Population

Parallel to Part III, Section V, the distribution of the hypothetical variances of a parent population which is assumed to be distributed according to Type III can be described. The fundamental relation of Theorems II and III hold:

$$\begin{aligned} \bar{\mu}_{2n:p} &= \bar{\mu}_{2n:z_y} \qquad \text{or} \qquad \alpha_{2n:p} = \alpha_{2n:z_y} \\ \bar{\mu}_{2n+1:p} &= -\bar{\mu}_{2n+1:z_y} \qquad\qquad \alpha_{2n+1:p} = -\alpha_{2n+1:z_y} \end{aligned}$$

But now $M_p = \dfrac{\sum (x - \hat{M}_x)^2}{r}$ (See Part IV, Section II)

$$(82) \qquad M_p = \frac{1}{r}\sum\left(x - m_1 + g\,\frac{\hat{\sigma}_x \hat{\alpha}_{3:x}}{1 + 2\hat{\delta}_{z_x}}\right)^2$$

$$= m_2 + g^2 \frac{\sigma_s^2 \hat{\alpha}_{3:x}^2}{(1 + 2\hat{\delta}_{z_x})^2 \phi^2(\hat{\alpha}_{3:x}, r, s)} \text{ [from (60)].}$$

[19] Salvosa, L. R., *Annals of Mathematical Statistics* Vol. I, No. II, 1930.

Upon the same assumption that the best approximated 'most probable values' of the mean, the standard deviation and the skewness be the mean, the standard deviation, and the skewness of the parent population, the distribution of $\bar{\mu}_{2:x}$ is characterized by

$$
(83)\quad \begin{cases}
M_{\bar{\mu}_{2:x}} = m_2 + g^2 \dfrac{\sigma_s^2 \hat{\alpha}_{3:x}^2}{(1+2\hat{\delta}_{s_x})^2 \phi^2(\hat{\alpha}_{3:x}, r, s)} = m_2 \left\{ 1 + \dfrac{g^2 \hat{\alpha}_{3:x}^2}{(1+2\hat{\delta}_{s_x})^2 \phi^2(\hat{\alpha}_{3:x}, r, s)} \right\} \\
\sigma_{\bar{\mu}_{2:x}} = \sigma_{s_y} = \sqrt{\dfrac{s-r}{r(s-1)}}\, \sigma_y = \sqrt{\dfrac{s-r}{r(s-1)}\,(3\gamma+2)}\, \sigma_x^2 \\
\qquad\qquad = \sqrt{\dfrac{s-r}{r(s-1)}\,(3\hat{\gamma}+2)}\, \dfrac{\sigma_s^2}{\phi^2(\hat{\alpha}_{3:x}, r, s)} \\
\alpha_{3:\bar{\mu}_{2:x}} = -\alpha_{3:s_y} = -\dfrac{s-2r}{s-2} \sqrt{\dfrac{s-1}{r(s-\gamma)}}\, \alpha_{3:y} \\
\qquad\qquad = -\dfrac{s-2r}{s-2} \sqrt{\dfrac{s-1}{r(s-r)}} \cdot \dfrac{30\hat{\gamma}^2 + 56\hat{\gamma} + 8}{(3\hat{\gamma}+2)^{\frac{1}{2}}} \\
\alpha_{4:\bar{\mu}_{2:x}} - 3 = \alpha_{4:s_y} - 3 = \dfrac{(s-1)(s^2+s-6rs)+6r^2)}{r(s-r)(s-2)(s-3)} [\alpha_{4:y} - 3] \\
\qquad\qquad - \dfrac{6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)} \\
\qquad\qquad = \dfrac{(s-1)(s^2+s-6rs+6r^2)}{r(s-r)(s-2)(s-3)} \left[\dfrac{630\hat{\gamma}^3 + 1680\hat{\gamma}^2 + 912\hat{\gamma} + 48}{(3\hat{\gamma}+2)^2} \right] \\
\qquad\qquad - \dfrac{6s(r-1)(s-r-1)}{r(s-r)(s-2)(s-3)}.
\end{cases}
$$

For an infinite parent population, we have

$$
(84)\quad \begin{cases}
M_{\bar{\mu}_{2:x}} = m_2 \left\{ 1 + \dfrac{1}{4r^2} \cdot \dfrac{\hat{\alpha}_{3:x}^2}{\phi^2(\hat{\alpha}_{3:x}, r)} \right\} \\
\sigma_{\bar{\mu}_{2:x}} = \sqrt{\dfrac{(3\hat{\gamma}+2)}{r}} \cdot \dfrac{m_2}{\phi^2(\hat{\alpha}_{3:x}, r)} \\
\alpha_{3:\bar{\mu}_{2:x}} = -\dfrac{1}{\sqrt{r}} \dfrac{30\hat{\gamma}^2 + 56\hat{\gamma} + 8}{(3\hat{\gamma}+2)^{3/2}} \\
\alpha_{4:\bar{\mu}_{2:x}} - 3 = \dfrac{1}{r} \left\{ \dfrac{630\hat{\gamma}^3 + 1680\hat{\gamma}^2 + 912\hat{\gamma} + 48}{(3\hat{\gamma}+2)^2} \right\}
\end{cases}
$$

Numerical illustration. Using the same sample in Table VIII, we wish to ascertain the probability that the variance of the parent population lies between

306.25 and 342.25. From $m_1 = 138.3$, $\sigma_s = 14.6366$, $\alpha_{3:s} = .81099$, and $\alpha_{4:s} = 4.47644$, we find from (84)

$$M_{\bar{\mu}_{2:x}} = 214.232,235$$

$$\sigma_{\bar{\mu}_{2:x}} = 34.335,74$$

$$\alpha_{3:\bar{\mu}_{2:x}} = -.495,311$$

$$\alpha_{4:\bar{\mu}_{2:x}} = 3.463,675,7$$

$$\delta_{s_y} = .105,515,6$$

From Part I, Section III,

$$k = \frac{\alpha^2_{3:\bar{\mu}_{2:x}}}{4\delta_{s_y}(2 + \delta_{s_y})} = .276 < 1$$

Therefore, the best fitting curve will be Type IV which assumes the form[20]

$$y = y_0 (1 + x^2)^{-m} e^{-\lambda \tan^{-1} x} \tag{85}$$

where

$$x = \frac{t + p}{q},$$

t being in standard units

$$p = \frac{b_1}{2b_2} = \frac{\alpha_3}{2\delta}$$

$$q^2 = \frac{4b_0b_2 - b_1^2}{4b_2^2} = \frac{4\delta(2 + \delta) - \alpha_3^2}{4\delta^2}$$

$$m = \frac{1}{2b_2} = \frac{1 + 2\delta}{\delta}$$

$$\lambda = -\frac{a + p}{b_2 q}$$

$$y_0 = \frac{e^{\frac{\lambda \pi}{2}}}{g(2m - 2, \lambda)} = \frac{1}{F(2m - 2, \lambda)}$$

y_0 is found from Pearson's *Tables for Statisticians and Biometricians*[21] to be .049662.

---

[20] Elderton, W. P., *op. cit.*, p. 64.

[21] Pearson, K., *Tables for Statisticians and Biometricians*, Vol. I, pp. 126–142.

Now the given limits 306.25 and 342.25 of the variance, when expressed in standard units, are

$$t_a = 2.679,941$$

$$t_b = 3.728,410$$

Therefore the probability that $\bar{\mu}_{2:x}$ lies between $\bar{\mu}_{2:x} = 306.25$ and $\bar{\mu}_{2:x} = 342.25$ is

$$P = y_0 \int_{t_a=2.679,041}^{t_b=3.728,410} (1 + x^2)^{-m} e^{-\lambda \tan^{-1}x} \, dx$$

we find

$$m = 11\ 477,271$$

$$\lambda = 12\ 940,307$$

$$P = .049662 \int_{08757}^{36343} (1 + x^2)^{-11.477271} e^{-12.940207 \tan^{-1}x} \, dx$$

By means of Maclaurin-Euler's Interpolation Formula, P is found to be equal to .000,904.

No definite law can be ascertained before we know $\hat{\alpha}_{3:x}$ because, as we have seen, $\alpha_{3:\bar{\mu}_{2:x}}$ and $\alpha_{4:\bar{\mu}_{2:x}}$ are both expressed in terms of s, r, and $\hat{\alpha}_{3:x}$. We do not know the value of k, which is a determining factor of the best fitting curve and a function of s, r, $\alpha_{3:\bar{\mu}_{2:x}}$ and $\alpha_{4:\bar{\mu}_{2:x}}$, until we know the values of s, r, and $\hat{\alpha}_{3:x}$.

Section VI. Distribution of the Hypothetical Third Moments oi the Parent Population About Its Mean

Recalling the fact that the distribution of the third moments of sample means about the most probable value of the mean of the parent population is equivalent to the consideration of a distribution of sample means drawn from a parent population, $w_1, w_2, w_3, \cdots w_s$ where $w_i = (x_i - \hat{M}_x)^3$, so we can write down in accordance with the fundamental relations stated in Theorems II and III:

$$\bar{\mu}_{2n:p} = \bar{\mu}_{2n:z_w} \qquad \alpha_{2n:p} = \alpha_{2n:z_w}$$

or

$$\bar{\mu}_{2n+1:p} = -\bar{\mu}_{2n+1:z_w} \qquad \alpha_{2n+1:p} = -\alpha_{2n+1:z_w}$$

But here $M_p = \dfrac{\sum (x - \hat{M}_x)^3}{r}$; and by the substitution of (60), we have

$$M_p = \frac{1}{r} \sum \left(x - m_1 + \frac{g \hat{\sigma}_x \hat{\alpha}_{3:x}}{1 + 2\hat{\delta}_{z_x}} \right)^3$$

(86)

$$= m_3 + 3m_2 \frac{g \sigma_s \hat{\alpha}_{3.x}}{(1 + 2\hat{\delta}_{z_x})\, \phi(\hat{\alpha}_{3:x}, r, s)} + \frac{g^3 \sigma_s^3 \hat{\alpha}_{3:x}^3}{(1 + 2\hat{\delta}_{z_x})^3 \phi^3(\hat{\alpha}_{3:x}, r, s)}$$

Consequently, with the same assumption that the best approximated 'most probable values' of the mean, the standard deviation, and the skewness be the mean, the standard deviation and the skewness of the parent population, the distribution of $\bar{\mu}_{3:x}$ is characterized by

$$
\begin{aligned}
M_{\bar{\mu}_{3:x}} &= m_3 + 3m_2 \frac{g\sigma_s \hat{\alpha}_{3:x}}{(1 + 2\hat{\delta}_{z_x})\phi(\hat{\alpha}_{3:x}, r, s)} + \frac{g^3 \sigma_S^3 \hat{\alpha}_{3:x}^3}{(1 + 2\hat{\delta}_{z_x})^3 \phi^3(\hat{\alpha}_{3:x}, r, s)} \\
\sigma_{\bar{\mu}_{3:x}} &= \sqrt{\frac{s - r}{r(s - 1)}}\, \sigma_{z_w} = \sqrt{\frac{s - r}{r(s - 1)}\,(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)}\, \hat{\sigma}_x^3 \\
&= \sqrt{\frac{s - r}{r(s - 1)}\,(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)}\, \frac{\sigma_S^3}{\phi^3(\hat{\alpha}_{3:x}, r, S)} \\
\alpha_{3:\bar{\mu}_{3:x}} &= -\frac{s - 2r}{s - 2}\sqrt{\frac{s - 1}{r(s - r)}}\, \alpha_{3:z_w} \\
&= -\frac{s - 2r}{s - 2}\sqrt{\frac{s - 1}{r(s - r)}}\, \frac{\hat{\alpha}_{3:x}(1215 + 6417\hat{\gamma} + 7938\hat{\gamma}^2 + 2520\hat{\gamma}^3)}{(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)^{\frac{3}{2}}}
\end{aligned}
\tag{87}
$$

$$
\begin{aligned}
\alpha_{4:\bar{\mu}_{3:x}} - 3 &= \frac{(s - 1)(s^2 + s - 6rs + 6r^2)}{r(s - r)(s - 2)(s - 3)}[\alpha_{4:z_w} - 3] - \frac{6s(r - 1)(s - r - 1)}{r(s - r)(s - 2)(s - 3)} \\
&= \frac{(s - 1)(s^2 + s - 6rs + 6r^2)}{r(s - r)(s - 2)(s - 3)} \cdot \frac{\begin{array}{r} 9720 + 417555\hat{\gamma} + 2707992\hat{\gamma}^2 + 5840343\hat{\gamma}^3 \\ + 4789530\hat{\gamma}^4 + 1247400\hat{\gamma}^5 \end{array}}{(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)^2} \\
&\quad - \frac{6s(r - 1)(s - r - 1)}{r(s - r)(s - 2)(s - 3)}.
\end{aligned}
$$

For an infinite parent population, we have

$$
\left\{
\begin{aligned}
M_{\bar{\mu}_{3:x}} &= m_3 + 3m_2 \frac{\sigma_s \hat{\alpha}_{3:x}}{2r\phi(\hat{\alpha}_{3:x}, r)} + \frac{1}{8r^3} \frac{\sigma_s^3 \hat{\alpha}_{3:x}^3}{\phi^3(\hat{\alpha}_{3:x}, r)} \\
\sigma_{\bar{\mu}_{3:x}} &= \sqrt{\frac{1}{r}\,(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)}\, \frac{\sigma_s^3}{\phi^3(\hat{\alpha}_{3:x}, r)} \\
\alpha_{3:\bar{\mu}_{3:x}} &= -\sqrt{\frac{1}{r}}\, \frac{\hat{\alpha}_{3:x}(1215 + 6417\hat{\gamma} + 7938\hat{\gamma}^2 + 2520\hat{\gamma}^3)}{(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)^{\frac{3}{2}}} \\
\alpha_{4:\bar{\mu}_{3:x}} - 3 &= \frac{1}{r} \frac{\begin{array}{r} 9720 + 417555\hat{\gamma} + 2707992\hat{\gamma}^2 + 5840343\hat{\gamma}^3 + 4789530\hat{\gamma}^4 \\ + 1247400\hat{\gamma}^5 \end{array}}{(15 + 63\hat{\gamma} + 30\hat{\gamma}^2)^2}
\end{aligned}
\right.
\tag{88}
$$

Numerical illustration. Using the same sample in Table VIII, we wish to ascertain the probability that the third moment of the parent population about

the mean lies between $\bar{\mu}_{3:s} = 3000$ and $\bar{\mu}_{3:s} = 4000$, still assuming an infinite parent population from which the sample is drawn

$$\hat{\alpha}_{3:s} = .6501$$
$$\phi(\hat{\alpha}_{3:s}, r) = 1.006,279$$

We find from (88)

$$M_{\bar{\mu}_{3:s}} = 2558.137,096$$
$$\sigma_{\bar{\mu}_{3:s}} = 1675.696,37$$
$$\alpha_{3:\bar{\mu}_{3:s}} = -1.187,409,9$$
$$\alpha_{4:\bar{\mu}_{3:s}} = 6.127,551,6$$
$$\delta_{s_w} = 0.221,886$$
$$k = 0.714,972 < 1$$

Therefore the best fitting curve is Type IV.

From Pearson's *Tables for Statisticians and Biometricians*, Vol. I,[22] we compute

$$y_0 = .000,058,032,3$$

The given limits 3000 and 4000 when expressed in standard units are $t = .263,689$ and $t = .860,455$ respectively. Therefore the probability that $\bar{\mu}_{3:s}$ lies between 3000 and 4000 may be expressed by

$$= y_0 \int_{t=.263689}^{t=.860455} (1 + x^2)^{-6.506819} e^{-17.443447 \tan^{-1} x}\, dx$$

By means of Maclaurin-Euler's Interpolation Formula, the answer is found to be .267,408,631.

We make the same remark here as we have made in the preceding two sections. That is, since $\alpha_{3:\bar{\mu}_{3:s}}$ and $\alpha_{4:\bar{\mu}_{3:s}}$ are both in terms of s, r and $\hat{\alpha}_{3:s}$,, we cannot determine the value of k which is a function of $\alpha_{3:\bar{\mu}_{3:s}}$ and $\alpha_{4:\bar{\mu}_{3:s}}$ until we know the values of s, r, and $\hat{\alpha}_{3:s}$. Consequently, the curve associated with the distribution of the hypothetical third moments of a parent population of Type III distribution is not known until we know s, r, and $\hat{\alpha}_{3:s}$.

[22] Pearson, K., *op. cit.*, pp. 126–142.

ON A METHOD OF TESTING THE HYPOTHESIS THAT AN OBSERVED SAMPLE OF n VARIABLES AND OF SIZE N HAS BEEN DRAWN FROM A SPECIFIED POPULATION OF THE SAME NUMBER OF VARIABLES

By John W. Fertig

With the Technical Assistance of Margaret V. Leary*

The problem of determining whether or not a given observation may be regarded as randomly drawn from a certain population completely specified with respect to its parameters is readily solved if the probability integral of that population be known. In particular if the population specified be a normal population, one may calculate the relative deviate $(x - a)/\sigma$, where a and σ are the population mean and standard deviation respectively, and refer to tables of the normal probability integral. The hypothesis that x was drawn from this population may be rejected if P is less than an arbitrarily fixed value, say $\leqq .01$. Generalizations of this problem may be made in two directions: 1) May a single observation simultaneously made on n variables be considered as randomly drawn from a specified population of n variables? 2) May a sample of one variable and of size N be regarded in its entirety as randomly drawn from a specified univariate population?

The solution to the first problem for the case of sampling from a normal population of n variables was given by Karl Pearson in 1908[1] as the "Generalized Probable Error." Let

$$\chi^2 = \frac{1}{P}\left\{ \sum_{i,j=1}^{n} P_{ij}\left[\frac{(x_i - a_i)(x_j - a_j)}{\sigma_i \sigma_j}\right]\right\}$$

where a_i and σ_i are the population mean and standard deviation respectively of the i^{th} variable, and P_{ij} is the usual cofactor of the element in the i^{th} row and j^{th} column of the determinant P of population correlation coefficients. That is,

$$P = |\rho_{ij}|\,;\ i, j = 1, 2, 3, \cdots, n.$$

The probability of an observation yielding a smaller discrepancy than that represented by the value of χ^2, i.e., lying between 0 and χ^2, may then be calculated from Tables of the Incomplete Normal Moment Functions[2]. The tables are entered in terms of $(\chi^2)^{\frac{1}{2}}$ and $(n - 1)$, and the tabled value multiplied by $(2\pi)^{\frac{1}{2}}$ or 2 depending upon whether n be even or odd respectively.

* From the Memorial Foundation for Neuro-Endocrine Research and the Research Service of the Worcester State Hospital, Worcester, Massachusetts.

The probability of an observation giving a greater discrepancy is then the complement of this value. Obviously, this latter probability may be obtained directly by entering tables of the X^2 distribution such as Elderton's[3] with n degrees of freedom, or through the use of Tables of the Incomplete Γ-Function[4].

The second problem, limited to the case of sampling from a normal population, was investigated by J. Neyman and E. S. Pearson in 1928[5]. The observed sample may be regarded as a point in N-dimensional space, where N is the sample size. Criteria for the acceptance or rejection of the hypothesis may be associated with contour surfaces in this space, so that in moving out from contour to contour the hypothesis becomes less and less reasonable. Frequently, contour surfaces on which the mean or standard deviation is constant are used for the testing of this hypothesis. Such surfaces are deficient inasmuch as they are not "closed" contours. Another contour system which appears more satisfactory is that of equiprobable pairs of m and s. The latter system in fact encloses roughly the same region as do the separate contours for the means and standard deviations. These systems are of course dependent on the particular statistics chosen to describe the sample and are further limited in that they do not take into account the probability of alternative hypotheses concerning the origin of the sample.

Using the principle of maximum likelihood Neyman and Pearson have developed a system of contours which is free of the above limitations. The system so derived is in fact quite similar to that of equiprobable pairs m and s. In a later paper[6], these same investigators have shown that this method of maximum likelihood does enable one to select the most efficient criteria for the testing of an hypothesis. The criterion selected on this basis is defined as

$$\lambda = \frac{\text{Likelihood that sample came from specified population}}{\text{Maximum likelihood that sample came from some other population}}$$

$$= (s^2/\sigma^2)^{N/2} e^{-N/2}\left[\frac{s^2 + (\bar{x} - a)^2}{\sigma^2} - 1\right]$$

where a and σ are the population mean and standard deviation respectively, and $\bar{x}$ and s the sample mean and standard deviation.

λ is constant upon certain contour surfaces in N-dimensional space, and diminishes on passing outward. The form of the surfaces is independent of N. It is evident that λ must lie between zero and unity. When it is close to unity we know that it is reasonable to assume that our hypothesis is true, when small we know that it is unreasonable. But we must know the probability of λ less than a certain value occurring when the hypothesis tested is true, so that we may control another source of error, namely, that of rejecting the hypothesis when it is true. In other words, we must know the sampling distribution of λ, so that we will reject the hypothesis only when the probability of obtaining a smaller value is negligible, say $P_\lambda \leqq .01$. Neyman and Pearson were not able to evaluate this distribution but they were able to integrate the original density function of the population appropriate to N-dimensional space outside of the

various λ contours. This they were able to do by effecting a transformation of the density function and contours to the plane of m and s. These values of P_λ have been tabled by them[7], the tables being entered in terms of N and k, where

$$k = \log\left[\frac{s^2 + (\bar{x} - a)^2}{\sigma^2}\right] - \log(s^2/\sigma^2)$$

The generalization of either of the above problems requires a criterion to test an hypothesis which may be formulated as follows: Given a sample Σ of n variables and of size N with means $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n$, standard deviations $s_1, s_2, \cdots, s_n$, and correlation coefficients $r_{12}, r_{13}, \cdots, r_{1n}, r_{23}, \cdots, r_{2n}, \cdots, r_{(n-1)n}$, may we regard this sample as randomly drawn from a population π of n variables and completely specified with respect to all its parameters? We shall restrict our inquiries to the case where π is a normal population. In this case the distribution law is

$$f(x_1, x_2, \cdots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2\cdots\sigma_n P^{\frac{1}{2}}} e^{-\phi}$$

where

$$\phi = +\frac{1}{2P}\left\{\sum_{i,j=1}^{n} P_{ij}\left[\frac{(x_i - a_i)(x_j - a_j)}{\sigma_i\sigma_j}\right]\right.$$

where a_i and σ_i are the population mean and standard deviation respectively of the i^{th} variable, and P and P_{ij} are as previously defined.

Thus the probability that Σ has been drawn from π with its N values of $x_{i\alpha}$ $(i = 1, 2, \cdots, n)$ lying in the interval $x_{i\alpha} \pm \frac{1}{2}dx_{i\alpha}$; $(\alpha = 1, 2, \cdots, N)$ is given by

$$C = \left[\frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2\cdots\sigma_n P^{\frac{1}{2}}}\right]^N e^{-\Theta}\, dX$$

where

$$\Theta = \frac{1}{2P}\left\{\sum_{i,j=1}^{n} P_{ij}\sum_{\alpha=1}^{N}\left[\frac{(x_{i\alpha} - a_i)(x_{j\alpha} - a_j)}{\sigma_i\sigma_j}\right]\right.$$

$$= \frac{N}{2P}\left\{\sum_{i,j=1}^{n} P_{ij}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a_i)(\bar{x}_j - a_j)}{\sigma_i\sigma_j}\right]\right.$$

$$dX = \prod_{i=1}^{n}\prod_{\alpha=1}^{N} dx_{i\alpha}$$

The likelihood that Σ has been drawn from any other normal population, such as π', is given by

$$C' = \left[\frac{1}{(2\pi)^{n/2}\sigma_1'\sigma_2'\cdots\sigma_n' P'^{\frac{1}{2}}}\right]^N e^{-\Theta'}\, dX$$

where

$$\Theta' = \frac{N}{2P'}\left\{\sum_{i,j=1}^{n} P'_{ij}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a'_i)(\bar{x}_j - a'_j)}{\sigma'_i \sigma'_j}\right]\right\}$$

The population from which it is most likely that Σ has been drawn is that for which C' is a maximum. The values of the parameters of this population may be obtained by putting

$$\frac{\partial C'}{\partial a'_i} = 0, \qquad \frac{\partial C'}{\partial \sigma'_i} = 0; \qquad (i = 1, 2, \cdots, n)$$

$$\frac{\partial C'}{\partial \rho'_{ij}} = 0; \qquad (i, j = 1, 2, \cdots, n)$$

These conditions are fulfilled when

$$a'_i = \bar{x}_i;\ \sigma'_i = s_i; \quad (i = 1, 2, \cdots, n)$$

$$\rho'_{ij} = r_{ij}; \quad (i, j = 1, 2, \cdots, n)$$

So that

$$C'_{\max.} = \left[\frac{1}{(2\pi)^{n/2} s_1 s_2 \cdots s_n R^{\frac{1}{2}}}\right]^N e^{-nN/2}$$

where

$$R = |r_{ij}|; \qquad i, j = 1, 2, \cdots, n$$

The appropriate criterion to select in order to test our hypothesis is thus

$$\lambda = \frac{C}{C'_{\max}} = \left[\frac{s_1 s_2 \cdots s_n R^{\frac{1}{2}}}{\sigma_1 \sigma_2 \cdots \sigma_n P^{\frac{1}{2}}}\right]^N e^{-w}$$

where

$$w = \frac{N}{2}\left\{\sum_{i,j=1}^{n} \frac{P_{ij}}{P}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a_i)(\bar{x}_j - a_j)}{\sigma_i \sigma_j}\right] - n\right\}$$

The equations λ = constant represent a series of contours in N-dimensional space. As we move outward from contour to contour our hypothesis becomes less and less acceptable. Although we may be confident that the use of this criterion will minimize the chance of accepting the hypothesis when it is false we must know the frequency with which samples occur outside of a given λ contour when the hypothesis is true. In other words, we must know the integral of C outside of various contours, or else we must know the sampling distribution of λ. The former is an exceedingly difficult method for n greater than unity. Thus for the case of $n = 2$ we should have to integrate some such expression as

$$k s_1^{N-2} s_2^{N-2} e^{-\Theta}(1 - r_{12}^2)^{\frac{N-4}{2}}\, d\bar{x}_1 d\bar{x}_2 ds_1 ds_2 dr_{12}$$

outside of the various contours. Nor have we so far been able to evaluate the sampling distribution. We can however give an expression for the moments of λ and thus reach an approximate distribution.

Wilks[8] has derived expressions for the moment coefficients about zero for the maximum likelihood criterion that k samples of n variables and of N_t observations each have been drawn from the same unspecified normal population of n variables. Thus,

$$\mu_h'(\lambda) = \prod_{t=1}^{n} \left[\frac{\sum_{t=1}^{k} N_t}{N_t} \right]^{\frac{h n N_t}{2}} \prod_{i=1}^{n} \left[\frac{\Gamma\left(\frac{N_t(1+h) - i}{2}\right)}{\Gamma\left(\frac{N_t - i}{2}\right)} \right]$$

$$\prod^{n} \left\{ \frac{\Gamma\left[\frac{\sum_{t=1}^{k} N_t - i}{2}\right]}{\Gamma\left[\frac{(1+h)\sum_{t=1}^{k} N_t - i}{2}\right]} \right\}$$

from which we can write expressions giving the moment coefficients about zero for the λ criterion for two samples

$$\mu_h'(\lambda) = \frac{(N_1 + N_2)^{\frac{n h (N_1+N_2)}{2}}}{N_1^{\frac{n h N_1}{2}} \; N_2^{\frac{n h N_2}{2}}}$$

$$\prod_{i=1}^{n} \left\{ \frac{\Gamma\left[\frac{N_1(1+h) - i}{2}\right] \Gamma\left[\frac{N_2(1+h) - i}{2}\right] \Gamma\left[\frac{N_1 + N_2 - i}{2}\right]}{\Gamma\left(\frac{N_1 - i}{2}\right) \Gamma\left(\frac{N_2 - i}{2}\right) \Gamma\left[\frac{(N_1 + N_2)(1+h) - i}{2}\right]} \right\}$$

The limit of this latter expression as $N_2 \to \infty$ will be the moment coefficient about zero for the λ criterion that one sample has been drawn from a specified population. Thus

$$\lim_{N_2 \to \infty} \mu_h'(\lambda) = \prod_{i=1}^{n} \frac{\Gamma\left[\frac{N_1(1+h) - i}{2}\right]}{\Gamma\left(\frac{N_1 - i}{2}\right)} \left(\frac{2e}{N_1}\right)^{\frac{n h N_1}{2}} \cdot (1+h)^{\frac{-n N_1 (1+h)}{2}}$$

Various roots of λ are distributed to a good degree of approximation according to a function of the form

$$f(t) = \frac{\Gamma(m_1 + m_2)}{\Gamma(m_1)\,\Gamma(m_2)} t^{m_1 - 1} (1 - t)^{m_2 - 1}$$

where

$$m_1 = \mu_1'(\mu_1' - \mu_2')/(\mu_2' - \mu_1'^2) ; \qquad m_2 = (1 - \mu_1')m_1/\mu_1'$$

and the value of μ_h' for roots of λ may be obtained by replacing h in the original expression by h times the desired root. Measures of the skewness and kurtosis of this distribution are given by

$$B_1 = 4(m_1 - m_2)^2(m_1 + m_2 + 1)/m_1m_2(m_1 + m_2 + 2)^2$$

$$B_2 = 3B_1(m_1 + m_2 + 2) + 6(m_1 + m_2 + 1)/2(m_1 + m_2 + 3)$$

A comparison with the true measures of skewness and kurtosis for various roots of λ as given by

$$B_1 = \mu_3^2/\mu_2^3 ; \qquad B_2 = \mu_4/\mu_2^2$$

will afford a measure of the goodness of the approximation and the range of values of N for which any particular root will be distributed as assumed.

Investigating the moments for n from one to four and N from three to fifty we note that in the case of samples of two and three variables, $\lambda^{1/N}$ follows the assumed distribution for N from 3 to 15; $\lambda^{2/N}$ from 15 to 30; $\lambda^{3/N}$ from 30 to 50. In the case of four variables, $\lambda^{1/2N}$ follows the distribution for N from 5 to 10; $\lambda^{1/N}$ from 10 to 20; $\lambda^{2/N}$ from 20 to 40; $\lambda^{3/N}$ from 40 to 50. It appears likely that for higher values of n, for N small, some such root as $\lambda^{1/2N}$ or $\lambda^{1/3N}$ will follow the assumed distribution, while as N increases smaller roots will follow it. For any value of n, the smallest permissible value of N is $(n + 1)$.

The probability that a smaller value of λ will be obtained when the sample has actually been drawn from π, i.e., P_λ, may thus be obtained by reference to Tables of the Incomplete B-Function[9] with $p = m_1$, $q = m_2$, $x =$ value of the particular root of the observed λ. We may also get the 1% and 5% levels of significance directly from Fisher's[10] tables of "z" or Snedecor's[11] tables of "F" $(= e^{2z})$, by taking

$$n_1 = 2m_2 ; \qquad n_2 = 2m_1 ; \qquad L = n_2/(n_2 + n_1F) ,$$

where L is the desired root of λ. Linear interpolation will generally suffice except for very small values of N.

For the case of $N \to \infty$, we have

$$\lim_{N\to\infty} \mu_h'(\lambda) = (1 + h)^{-\sum_{i=2}^{n+1} i/2}$$

Thus the quantity $(-2 \log \lambda)$ will be distributed in the χ^2 distribution with $\sum_{i=2}^{n} i$ degrees of freedom.

A table of the 1% and 5% levels of significance for n equal one to four, and values of N from five to ∞ is given below

5% and 1% Levels of Significance of "λ"

— N —

n		5	10	15	20	30	40	50	∞
1	5%	.025	.037	.041	.043	.045	.046	.047	.050
	1%	.003	.006	.008	.008	.009	.009	.009	.010
2	5% × 10^{-2}	.046	.173	.234	.269	.308	.330	.343	.392
	1% × 10^{-3}	.026	.168	.260	.305	.372	.409	.428	.525
3	5% × 10^{-3}	.001	.036	.072	.097	.125	.143	.155	.211
	1% × 10^{-4}	.000+	.019	.047	.076	.101	.117	.128	.194
4	5% × 10^{-5}		.026	.106	.174	.295	.356	.418	.710
	1% × 10^{-6}		007	.040	.075	.145	.185	.221	.466

A check on the accuracy of the method of approximation used may be obtained by comparing the values of P_λ for the case of $n = 1$ with the exact values given by Neyman and Pearson. For $n = 10$, $\lambda^{1/N}$ is distributed as assumed with $m_1 = 9.0562$, $m_2 = 0.9987$. For the case of $(\bar{x} - a)/\sigma = 0.2$, $s/\sigma = 1.2$, we find $k = 0.48439$, $\lambda^{1/N} = .94395$. From the Tables of the Incomplete B-Function we find $P_\lambda = .5936$, from Neyman and Pearson's tables, .5935.

No studies have been made on the extent of deviation from normality permissible for the application of the test. There is no reason to doubt, however, that as much deviation is permissible as in the case of the univariate λ. From theoretical considerations and from sampling studies Neyman and Pearson conclude that the univariate λ technique holds for deviation from normality to the extent of ± 0.5 for B_1 and 2.5 to 4.2 for B_2.

We are confident that this generalized λ technique will be found useful in biological research. If the n variables were uncorrelated we would be able to test whether the sample had been drawn from the population of n variables by successive applications of the univariate λ technique and then combining the resulting probabilities. In general, however, there will be some correlation between the variables, however slight. The method here proposed will take account of all possible intercorrelations, and consequently all multiple and partial correlations.

Now, if P_λ is less than some arbitrarily fixed value, say $\leqq .01$, we may decide which variable or variables contributes most to this result, by performing simpler λ tests. It may be due to one or more of the means, standard deviations,

or correlation coefficients. As may often be the case, it is not due to any one factor but to contributions from all of them. That is, all possible factors tested separately might show a fairly reasonable value of P, but if all the separate values are combined somehow, as by means of this λ method, the resultant P may be too small. It is in such problems that this technique should provide valuable information.

In case k samples of n variables are available it should be possible to determine whether all of them have come from the same specified population of n variables by performing k λ tests and combining the separate values of P_λ. Such a hypothesis may best be tested, however, by a further extension of the λ theory which the writers are at present investigating.

The following problem is chosen to illustrate the computations involved in the application of the test. Many of the investigations pursued at the Worcester State Hospital attempt to differentiate between schizophrenic patients and normal controls. In one such type of investigation various blood constituents were determined, namely, Urea N_2 (mg./100 cc.), Uric Acid N_2 (mg./100 cc.), Creatine N_2 (mg./100 cc.) for a sample of twenty-five schizophrenic patients. Previous investigations on these same variables for a large series of normal controls yielded constants which for the purpose of the example may be considered as the population parameters. Past studies on these variables have not shown any marked degree of non-normality for the various distributions.

These variables are designated as

$$1 = \text{Urea } N_2\,; \qquad 2 = \text{Uric Acid } N_2\,; \qquad 3 = \text{Creatine } N_2$$

The parameters of the population are given by

$$a_1 = 16.03\,; \qquad a_2 = 1.40\,; \qquad a_3 = 1.25$$

$$\sigma_1^2 = 20.268\,; \qquad \sigma_2^2 = 0.029\,; \qquad \sigma_3^2 = 0.025$$

$$\rho_{12} = .3075\,; \qquad \rho_{13} = .1232\,; \qquad \rho_{23} = .3853$$

The statistics for the sample of twenty-five are

$$\bar{x}_1 = 15.56\,; \qquad \bar{x}_2 = 1.42\,; \qquad \bar{x}_3 = 1.25$$

$$s_1^2 = 10.486\,; \qquad s_2^2 = 0.043\,; \qquad s_3^2 = 0.025$$

$$r_{12} = -.0161\,; \qquad r_{13} = .0925\,; \qquad r_{23} = .2174$$

None of these statistics differs significantly from the corresponding parameters.

$$R = 0.9443\,; \qquad P = 0.7710\,;$$

$$P_{12}/P = -0.3373\,; \qquad P_{13}/P = -0.0061\,; \qquad P_{23}/P = -0.4506\,;$$

$$P_{11}/P = 1.1045\,; \qquad P_{22}/P = 1.2773\,; \qquad P_{33}/P = 1.1744$$

$$w = 12.5\ (0.3802) = 4.7531$$

$$\cdot (s_1^2\ s_2^2\ s_3^2\ R/\sigma_1^2\ \sigma_2^2\ \sigma_3^2\ P) = 0.9001$$

$$\log \lambda = 12.5 \log (0.9001) - 4.7531 \log e = \bar{3}.3641$$

$$\lambda = .0023$$

Since the 5% level of significance is about .0001, we thus conclude that the patients are not differentiated from the control population with respect to these variables.

REFERENCES

1. Pearson, Karl. Biometrika, vol. 6, 1908. pp. 59–68.
2. Tables for Statisticians and Biometricians, Part I. pp. xxiv–xxviii, 22–23.
3. Ibid. pp. xxxi–xxxiii, 26–28.
4. Tables of the Incomplete Γ-Function, 1934.
5. Neyman, J. and Pearson, E. S. Biometrika, vol. 20, 1928. pp. 175–241.
6. Ibid. Phil. Trans. Roy. Soc. A, vol. 231, 1933. pp. 289–337.
7. Tables for Statisticians and Biometricians, Part II. pp. clxxx–clxxxv, 221–223.
8. Wilks, S. S. Biometrika, vol. 24, 1932. pp. 471–494.
9. Tables of the Incomplete Beta-Function, 1934.
10. Fisher, R. A. Statistical Methods for Research Workers. Fourth Edition, 1932.
11. Snedecor, G. W. Calculation and Interpretation of Analysis of Variance and Covariance, 1934.

ON CONFIDENCE RANGES FOR THE MEDIAN AND OTHER EXPECTATION DISTRIBUTIONS FOR POPULATIONS OF UNKNOWN DISTRIBUTION FORM

By William R. Thompson

About the commonest situation with which we are confronted in mathematical statistics is that where we have a sample of n observations, $\{x_i\}$, which is assumed to have been drawn at random from an unknown population, U, with a zero probability that any two values in the finite sample be equal; and we desire to obtain from this evidence some insight as to parameters of the parent population, U. If further assumptions are made as to some of the parameters or the form of U, there may result a gain in power in testing other given hypotheses or establishing *confidence ranges* for particular parameters, but at an obvious sacrifice of scope in application. Insistent problems involve estimation of mathematical expectation that in further sampling we shall find x lying within a given interval, or similar expectation with regard to parameters of U such as the unknown median. It might seem that, without further assumption, all we should claim is that it is possible to draw from U the sample actually observed. A mere description of the experience may well be considered the observer's first duty, but a restriction to this would leave entirely unused the quality of *randomness* which has been assumed. What additional statements as to U may be appropriate in view of this randomness are our immediate concern; and the object of the present communication is to show how we may obtain such expressions in the form of mathematical expectations, and to present some results. Widespread applications to problems of estimation of *normal ranges of variation* or specific confidence ranges and comparisons of sample reflections of possibly different populations are immediately suggested, and a new foundation is offered for the study of frequency-distribution from the point of view of Schmidt.[1]

Section 1

Accordingly, consider the following situation. Let $A = \{x\}$ denote the set of all real numbers; and U denote an unknown frequency-distribution law of draft from $\{x\}$ such that there exists an unknown function, $f(x)$, bounded and not negative in A, and that the probability of obtaining x in an arbitrary interval (α, β) is

$$P(\alpha < x < \beta) = \int_{\alpha}^{\beta} f(x) \cdot dx \; ; \tag{1}$$

[1] Schmidt, R., *Annals of Math. Stat.*, 5, 30, (1934).

and, for every positive $p < 1$, there exists a finite interval (α, β) such that $P(\alpha < x < \beta) > p$. Let U be called an *infinite population*; and let n drafts, independently thus governed, made from A *without replacements* be called *a random sample of n observations from U*. Let $S = \{x_k\}$, $k = 1, \cdots, n$, denote such a sample; the enumeration to be made in an arbitrarily determined manner. In any case $x_i \neq x_j$ for $i \neq j$.

Temporarily, let us consider k to indicate the order of draft of the values of $\{x_k\}$, and let $p_k = P(x < x_k)$ denote the probability that x, drawn at random from U, be less than x_k of S. The probability *à priori* (i.e., without regard to relative values of x in the sample) that in such random sampling p_k lie between p' and p'', where $0 \leqq p' < p'' \leqq 1$, is obviously independent of k, and equals $p'' - p'$; i.e., p_k is equally likely *à priori* to lie in either of any two equal intervals in its possible range, (0, 1). Furthermore, the probability that in the rest of the sample, S, there will be just r values less than x_k is

$$\binom{n-1}{r} \cdot p_k^r \cdot (1 - p_k)^n$$

where r is an integer and $0 \leqq r < n$. Of course, p_k is unknown; but we may calculate (for all cases in repeated sampling wherein the same value of r is encountered) the expectation, $\overline{P}_r$ $(p' < p_k < p'')$, that p_k lie in the interval (p', p''). This is given by

$$\overline{P}_r(p' < p_k < p'') = \frac{(r+s+1)!}{r!\,s!} \int_{p'}^{p''} p^r \cdot q^s \cdot dp\,, \tag{2}$$

where $s = n - 1 - r$, and $q = 1 - p$. This is a familiar result[2,3,4] in applications of the well-known principle of Bayes to estimation of *à posteriori* probability. The approach is convenient in that many relations which have been developed in this connection are made immediately available. However, that p_k is equally likely *à priori* to lie in either of any two equal intervals in its possible range, is not based in the present case upon an especially added assumption nor any plea concerning *equal distribution of ignorance*, but follows directly from the elementary assumptions of random sampling. Accordingly, we are enabled to develop for given ranges what may be called the *specific confidence* or mathematical expectation that a given variable lie therein.

Obviously, (2) does not depend on k if this index is the order of draft provided that just r values of the sample, S, are less than the one under consideration, x_k. To simplify notation, accordingly, let the index k for any given sample, $\{x_k\}$,

[2] Bayes, *Philosophical Transactions*, *53*, 370 (1763). *Cf.* Todhunter, I., "A History of the Mathematical Theory of Probability," Macmillan and Co., London, 1865.

[3] Laplace, "Théorie Analytique des Probabilités," Paris, 1820; and other works, *Cf.* Todhunter, *l.c.*

[4] Pearson, K., *Philosophical Magazine*, *Series 6*, *Vol. 13*, 365, (1907).

be determined by the relations, $x_i < x_j$ for $i < j$, where $k = 1, \cdots, n$. Then, by (2) as $k = r + 1$, we have

$$\bar{P}(p' < p_k < p'') = \frac{n!}{(k-1)!\,(n-k)!} \int_{p'}^{p''} p^{k-1} \cdot q^{n-k} \cdot dp\,, \tag{3}$$

where p_k is the probability that random sample values from U will be less than the k-th value in order of ascending magnitude from a given random sample, $\{x_k\}$, of n values from U; and $\bar{P}(p' < p_k < p'')$ denotes the expectation that in such sampling p_k will lie in the interval, (p', p'').

In general, let $E(w) \equiv \bar{w}$ denote the mathematical expectation of any variable, w, under the given sampling conditions. Then, from a well-known relation developed by Laplace, we obtain from (3) the mean expectation of p_k,

$$\bar{p}_k = \frac{k}{n+1}\,; \tag{4}$$

and, further relations[4] of Karl Pearson yield

$$E((p_k - \bar{p}_k)^2) = \overline{\sigma^2_{p_k}} = \frac{k(n-k+1)}{(n+1)^2 \cdot (n+2)}\,; \tag{5}$$

i.e., the mean squared error in systematic use of $\frac{k}{n+1}$ instead of the unknown p_k should have the value in (5). Specific confidence ranges for x are readily established; e.g., the expectation that in random draft from U we obtain x within the range (x_k, x_{n-k+1}) in view of the sample, S, is

$$\bar{P}(x_k < x < x_{n-k+1}) = \frac{n+1-2k}{n+1}, \quad \text{for } 2k < n+1\,; \tag{6}$$

and $\bar{P}(x < x_k) = \bar{P}(x > x_{n-k+1}) = \frac{k}{n+1}$. For a given variate, w, the range (α, β) will be called *central* if $\bar{P}(w < \alpha) = \bar{P}(w > \beta)$, as in the case under (6). This is in accord with the development of the subject of confidence ranges by Neyman[5,6] and by Clopper and E. S. Pearson[7] following the introduction of the notion of fiducial interval by R. A. Fisher.[8,9] The estimates of p_k in (4) may be of value in studying frequency-distribution from the point of view developed by Schmidt,[1] by comparison of x_k with $\psi\left(\frac{k}{n+1}\right)$ rather than $\psi\left(\frac{2k-1}{2n}\right)$ where ψ is a univariant inverse of the integral of a given frequency function, taken to

[5] Neyman, J., *J. Roy. Stat. Soc.*, *97*, 589, (1934).
[6] Neyman, J., *Annals of Math. Stat.*, *6*, No. 3, 111, (1935).
[7] Clopper, C. J., and Pearson, E. S., *Biometrika*, *26*, 404, (1934).
[8] Fisher, R. A., *Proc. Camb. Phil. Soc.*, *26*, 528, (1930).
[9] Fisher, R. A., *Proc. Roy. Soc.*, *A 139*, 343, (1933).

replace the unknown $f(x)$. Obviously, $\bar{P}(x_k < x < x_{k+1}) = \dfrac{1}{n+1}$. A discussion of the special case, $n = 2$, has been prominent recently in a controversy between Jeffrey[10] and Fisher[9,11] and in an article by Bartlett.[12]

Now, in (3) for $p = p'$, and $p'' = 1$; we may write[13]

$$(7) \quad \bar{P}(p < p_k) = \sum_{\alpha=0}^{k-1} \binom{n}{\alpha} \cdot p^{\alpha} \cdot q^{n-\alpha} \equiv I_q(n-k+1, k) \equiv \frac{B_q(n-k+1, k)}{B_1(n-k+1, k)},$$

where $q = 1 - p$, and the incomplete B and I functions are those of K. Pearson[13] and Müller.[14] Now, let M be the unknown median of the infinite population, U. Then, by definition of p_k, if and only if $x_k > M$, then $p_k > \frac{1}{2}$. Therefore,

$$(8) \quad \bar{P}(M < x_k) = \bar{P}(0.5 < p_k) = \left(\frac{1}{2}\right)^n \cdot \sum_{\alpha=0}^{k-1} \binom{n}{\alpha} \equiv I_{0.5}(n-k+1, k).$$

Obviously, $\bar{P}(x_k < M < x_{k+1}) = \left(\frac{1}{2}\right)^n \cdot \binom{n}{k}$, and the expectation that M lie between the k-th observations from each end of the set, S, is given by

$$(9) \quad \bar{P}(x_k < M < x_{n-k+1}) = 1 - 2 \cdot I_{0.5}(n-k+1, k), \text{ for } 2k < n+1.$$

Obviously, this confidence range is *central.*

Section 2

Now, consider another infinite population, U'. In similar manner we may develop expressions for confidence ranges and distribution expectations. Let x' be the variate, and consider a sample, $S' = \{x'_m\}$, of n' observations drawn *without replacements* from A according to U' but after the sample, S, of U; i.e., so that no two of these sample values in S' are equal, nor any of them equal to a value in S. Furthermore, let m be the order of ascending magnitude of x' values in S'; and $p'_m \equiv P(x' < x'_m)$ for x' drawn at random from U', and let M' be the unknown median of U'. Then, by replacement of x, n, p_k, k, and M by x', n', p'_m, m, and M', respectively, in relations already developed for U and S, we obtain corresponding expressions for U' and S'; e.g.,

$$(10) \qquad \bar{P}(x'_m < x' < x'_{m+1}) = \frac{1}{n'+1}.$$

[10] Jeffreys, H., *Proc. Roy. Soc., A 138*, 48, (1932); *A 140*, 523, (1933); *A 146*, 9, (1934); *Proc. Camb. Phil. Soc., 29*, 83, (1933).

[11] Fisher, R. A., *Proc. Roy. Soc., A 146*, 1, (1934).

[12] Bartlett, M. S., *Proc. Roy. Soc., A 141*, 518, (1933).

[13] Pearson, K., *Biometrika, 16*, 202, (1924).

[14] Müller, J. H., *Biometrika, 22*, 284, (1930–31).

Now, let the index values, k_m, be defined as the number of values of $\{x_k\}$ that are less than x'_m, $m = 1, \cdots, n'$. Then, for all realized cases,

$$x_{k_m} < x'_m < x_{k_m+1}, \qquad m = 1, \cdots, n', \tag{11}$$

for the extreme members of (11) in S. Then, for x and x' drawn at random from U and U', respectively, we may write

$$0 < (n+1)(n'+1)\cdot\bar{P}(x < x') - \sum_{m=1}^{n'} k_m < n + n' + 1, \tag{12}$$

provided that the expectations for U and U' may be treated as independent. Similarly, for $\bar{P}(M < M')$ we have the relations,

$$\sum_{m=1}^{n'} \binom{n'}{m}\cdot I_{0.5}(n - k_m + 1, k_m) < 2^{n'}\cdot\bar{P}(M < M') < 1 + \sum_{m=1}^{n'} \binom{n'}{m-1}\cdot I_{0.5}(n - k_m, k_m + 1). \tag{13}$$

Of course, $I_{0.5}(n+1, 0) \equiv 0$, and $I_{0.5}(0, n+1) \equiv 1$. It may be verified readily that the inequality relations of (12) and (13) provide *best* upper and lower bounds for $\bar{P}(x < x')$ and $\bar{P}(M < M')$ under the circumstances given.

Obviously, any increasing function, $\phi(y)$, for y in A, may be used throughout the arguments, with $\phi(y)$ replacing $y = x, x_k, M, x', x'_m, M'$, respectively.

Section 3

Consider, now, the case of a finite population, U_N, of real numbers $\{x^{(i)}\}$, $x^{(i)} < x^{(j)}$ for $i < j$, $i = 1, \cdots, N$. Assume that N is known, and that a sample, S, of n values has been drawn at random from U_N without replacements. Let the sample values be $\{x_k\}$, $k = 1, \cdots, n$; and k be an arbitrarily determined index. As before, we might consider k the order of draft, temporarily, but the same analysis may be made if we let k be the order of ascending magnitude in the sample, S, and disregard its value in connection with *à priori* estimates of draft probability. Each $x_k = x^{(u_k)}$ for some unknown $u_k = 1, \cdots, N$; and, *à priori* (i.e., with no knowledge as to order of magnitude of other values in the sample), any two of these values are equally likely. Obviously, this is so if x_k is the first value drawn from U_N, and the rest of the sample may be regarded as a random draft without replacements of $n - 1$ elements from $[U_N - x_k]$. Let r be the number of these sample values less than x_k, and $s = n - 1 - r$. Then the probability of drawing such a sample after the given x_k, under the conditions given, is $\dfrac{\binom{u_k - 1}{r}\binom{N - u_k}{s}}{\binom{N-1}{n-1}}$, where $u_k - 1$ is the unknown number of

values in U_N that are less than x_k. To estimate the expectation, $\overline{P}(R = u_k - 1)$, that there are just a given number, R, of values in U_N less than x_k; we encounter the same situation considered by K. Pearson in a paper[15] subsequent to those applied to the infinite universe; and, by a simple conversion in notation, we have

$$\overline{P}(R = u_k - 1) = \frac{\binom{R}{r}\cdot\binom{N-1-R}{s}}{\binom{N}{n}}. \tag{14}$$

In previous communications[16,17] I have defined a function,

$$\psi(r, s, r', s') \equiv \frac{\sum_{\alpha=0}^{r'}\binom{r+r'-\alpha}{r}\cdot\binom{s+s'+1+\alpha}{s}}{\binom{r+s+r'+s'+2}{r+s+1}}, \tag{15}$$

for any four rational integers r, s, r', $s' \geqq 0$; and shown that Pearsons further result, equivalent here to evaluation of $\overline{P}(u_k \leqq R + 1)$ for a given R, may be expressed by means of this ψ-function. Thus, we have

$$\overline{P}(u_k \leqq R + 1) = \psi(r, s, R - r, N - R - s - 2). \tag{16}$$

It was demonstrated also[16 17] that

$$\psi(r, s, r', s') \equiv \psi(r, r', s, s') \equiv \psi(s', r', s, r) \equiv 1 - \psi(s, r, s', r') \tag{17}$$

with extension of the definition to include $\psi(r, s, -1, s') \equiv 0$, and that

$$\psi(r, s, r', s') \equiv \frac{\sum_{\alpha=0}^{\alpha' \leqq s, r'}\binom{r+r'+1}{r+1+\alpha}\cdot\binom{s+s'+1}{s-\alpha}}{\binom{r+s+r'+s'+2}{r+s+1}}. \tag{18}$$

As in the case of the infinite population, here also it is obvious that the order of draft of x_k is of no consequence in the analysis; and again we will let $k = r + 1$, whence $s = n - k$, and we may make these substitutions in (14) and (16). Then, we may write

$$\overline{P}(u_k \leqq R) = \psi(k - 1, n - k, R - k, k + N - R - n - 1); \tag{19}$$

---

[15] Pearson, K., *Biometrika*, *20 A*, 149, (1928).
[16] Thompson, W. R., *Biometrika*, *25*, 285, (1933).
[17] Thompson, W. R., *American Journal of Mathematics*, *57*, 450, (1935).

and, obviously, $\bar{P}(u_{n-k+1} \geqq N - R + 1) \equiv \bar{P}(u_k \leqq R)$. Hence, if we let M be the unknown median of U_N; and $m \equiv \frac{N-a}{2}$, where $a = 0, 1$, and $N - a$ is even; then, as u_i is an integer,

$$\bar{P}(x_k \leqq M \leqq x_{n-k+1}) \equiv \bar{P}\left(u_k \leqq \frac{N}{2} \leqq u_{n-k+1}\right) \tag{20}$$
$$\equiv 1 - 2\cdot\psi(k-1, n-k, m-k, k+N-m-n-1),$$

which is the expectation that the median of U_N lie within the closed interval, (x_k, x_{n-k+1}), for $2k \leqq n + 1$. This gives the confidence range, analogous to that for the infinite universe. It may be noted that

$$\bar{P}(u_k \leqq R < u_{k+1}) = \bar{P}(u_k \leqq R) - \bar{P}(u_{k+1} \leqq R)$$
$$= \psi(r, s, r', s') - \psi(r+1, s-1, r'-1, s'+1)$$

where $r = k - 1$, $s = n - k$, $r' = R - k$, and $s' = k + N - R - n - 1$. Hence, (18) gives

$$\bar{P}(u_k \leqq R < u_{k+1}) \equiv \frac{\binom{R}{k}\cdot\binom{N-R}{n-k}}{\binom{N}{n}}. \tag{21}$$

The approach by way of Pearson's problem again makes it easy to evaluate the expected mean p_k and variance as in the case of the infinite population, where $p_k = P(x < x_k)$ for x drawn at random from U_N. Of course, $p_k = \frac{u_k - 1}{N}$, but u_k is unknown. From Pearson's result,[15] however, we obtain

$$\bar{p}_k = \frac{k(N+1) - n - 1}{N(n+1)} = \frac{k}{n+1}\left(1 - \frac{n}{N}\right) + \frac{k-1}{N}, \tag{22}$$

and the expected variance of p_k,

$$\overline{\sigma^2_{p_k}} = E((p_k - \bar{p}_k)^2) = \frac{k(n-k+1)(N+1)(N-n)}{(n+1)^2\cdot(n+2)\cdot N^2}. \tag{23}$$

YALE UNIVERSITY.

THE SAMPLING DISTRIBUTION OF THE COEFFICIENT OF VARIATION

By Walter A. Hendricks with the assistance of Kate W. Robey

National Agricultural Research Center, Beltsville, Maryland

The coefficient of variation does not appear to be of very great interest to statisticians in general. However, its use in biometry is sufficiently extensive for some knowledge of its sampling distribution to be desirable. The present paper is an attempt to satisfy this need.

For the purposes of the following discussion, the coefficient of variation may be defined as the ratio of the standard deviation of a number of measurements to the arithmetic mean:

$$v = \frac{s}{\bar{x}} \qquad (1)$$

As is well known, the probability that the mean of a sample of n measurements, taken at random from a normal universe, lies between $\bar{x}$ and $\bar{x} + d\bar{x}$ and that the standard deviation of the measurements in the same sample lies between s and $s + ds$ is given by the relation:

$$dF_{\bar{x},s} = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\, e^{-\frac{n}{2\sigma^2}[(\bar{x}-m)^2+s^2]}\, s^{n-2}\, d\bar{x}\, ds \qquad (2)$$

If equation (2) is expressed in terms of polar coördinates by means of the transformation: $\bar{x} = \rho \cos\theta$; $s = \rho \sin\theta$, it becomes a distribution function of ρ and θ in which $\theta = \arctan v$:

$$dF_{\rho,\theta} = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\, e^{-\frac{n}{2\sigma^2}(\rho^2-2m\rho\cos\theta+m^2)}\, \rho^{n-1} \sin^{n-2}\theta\, d\rho\, d\theta \qquad (3)$$

In equation (3), ρ may vary from 0 to ∞ and θ may vary from 0 to π. To find the distribution function of θ, all that is necessary is to write:

$$dF_\theta = k\left[\int_0^\infty e^{-(a\rho-b)^2}\, \rho^{n-1}\, d\rho\right] d\theta \qquad (4)$$

in which,

$$k = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\, e^{-\frac{n}{2\sigma^2}m^2\sin^2\theta}\sin^{n-2}\theta,$$

$$a = \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\,\sigma}, \quad \text{and} \quad b = \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\,\sigma}\, m\cos\theta,$$

and to perform the indicated integration.

To evaluate the integral inside the brackets in equation (4), we may write:

$$\int_0^\infty e^{-(a\rho-b)^2}\rho^{n-1}\,d\rho = \frac{1}{a^n}\int_{-b}^\infty e^{-u^2}\sum_{i=0}^{n-1}\frac{(n-1)!}{(n-1-i)!\,i!}\,u^{n-1-i}\,b^i\,du. \quad \ldots(5)$$

Consider the integral, $\int_{-b}^\infty e^{-u^2}\,u^{n-1-i}\,du$. If b is sufficiently large, as is the case when the parameters of equation (2) are of such magnitude that practically the entire volume under the frequency surface lies to the right of the s axis, that is to say, if negative and small positive values of $\bar{x}$ occur so infrequently that their effects may be neglected, the lower limit, $-b$, of this integral may be replaced by $-\infty$ without introducing any appreciable error. The value of the integral, $\int_{-\infty}^\infty e^{-u^2}\,u^{n-1-i}\,du$, is zero when $n-1-i$ is odd and $\Gamma\left(\frac{n-i}{2}\right)$ when $n-1-i$ is even, zero being counted as an even number.

Subject to the above condition that b be sufficiently large, we may, therefore, write equation (5) in the form:

$$\int_0^\infty e^{-(a\rho-b)^2}\rho^{n-1}\,d\rho = \frac{1}{a^n}\sum_{i=0}^{n-1}{}'\frac{(n-1)!}{(n-1-i)!\,i!}\,\Gamma\left(\frac{n-i}{2}\right)b^i. \qquad .(6)$$

in which the symbol, $\sum'$, indicates that the only terms entering into the summation are those in which $n-1-i$ is an even number.

Substituting this expression for the integral inside the brackets in equation (4), replacing k, a, and b by the quantities which they represent, and writing V in place of the ratio, $\frac{\sigma}{m}$, we obtain the following distribution function of θ:

$$dF_\theta = \frac{2}{\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}\, e^{-\frac{n}{2V^2}\sin^2\theta}\sin^{n-2}\theta\sum_{i=0}^{n-1}{}'\frac{(n-1)!\,\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!\,i!}\,\frac{n^{\frac{1}{2}i}}{2^{\frac{1}{2}i}\,V^i}\cos^i\theta\,d\theta. \quad (7)$$

Equation (7) may be written in terms of v, if desired, by making the substitution, $\theta = \arctan v$:

$$dF_v = \frac{2}{\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{n}{2V^2}\frac{v^2}{1+v^2}} \frac{v^{n-2}}{(1+v^2)^{\frac{1}{2}n}}$$

$$\sum_{i=0}^{n-1}{}' \frac{(n-1)!\,\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!\,i!} \frac{n^{\frac{1}{2}i}}{2^{\frac{1}{2}i}V^i} \frac{1}{(1+v^2)^{\frac{1}{2}i}}\,dv \quad . . (8)$$

It must be emphasized that equation (8) has been derived on the hypothesis that negative and small positive values of $\bar{x}$ occur so infrequently that they may be neglected. However, since this condition is satisfied in the vast majority of practical problems in which the coefficient of variation is likely to be used, the limitation is not of much practical importance.

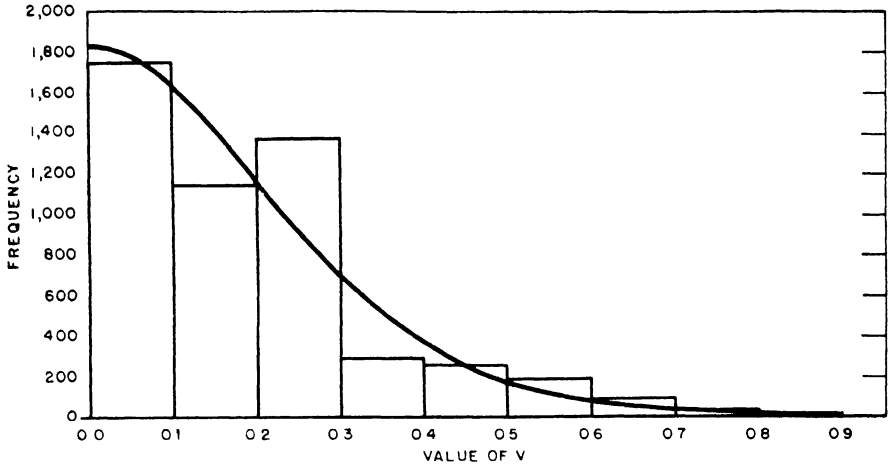


FIG. 1. OBSERVED AND THEORETICAL DISTRIBUTIONS OF VALUES OF v FOR 512 SAMPLES OF NUMBERS OF HEADS APPEARING IN TWO SUCCESSIVE TOSSES OF TEN COINS

As a test of the validity of equation (8), the authors calculated 512 coefficients of variation of the numbers of heads appearing in two successive tosses of ten coins. The coins were tossed 1024 times, thus yielding 512 samples, each consisting of two observations. For these data we have $m = 5$, $\sigma = 1.581$, and $V = 0.3162$.

For the case, $n = 2$, equation (8) reduces to:

$$dF_v = \frac{2}{\pi^{\frac{1}{2}}V} e^{-\frac{1}{V^2}\frac{v^2}{1+v^2}} \frac{dv}{(1+v^2)^{\frac{1}{2}}} \quad . . . . . (9)$$

Figure 1 shows the distribution of the 512 values of v obtained from the coin tossing experiment, together with the theoretical distribution given by equation (9).

An inspection of Figure 1 indicates that the agreement between the observed

and theoretical frequencies is fairly good. An application of the familiar chi test for goodness of fit showed the agreement to be rather poor. According to this test, the degree of discrepancy between theory and observation could have arisen by chance less than once in a hundred trials. However, the discrepancies may be partly due to the fact that data distributed in a discrete fashion were treated by methods appropriate to the analysis of data distributed according to a continuous frequency curve.

As another test of the validity of equation (8), the authors calculated 149 coefficients of variation of "days to maturity," which is the length of time elapsing between the date of hatch of a chicken and the time egg production com-

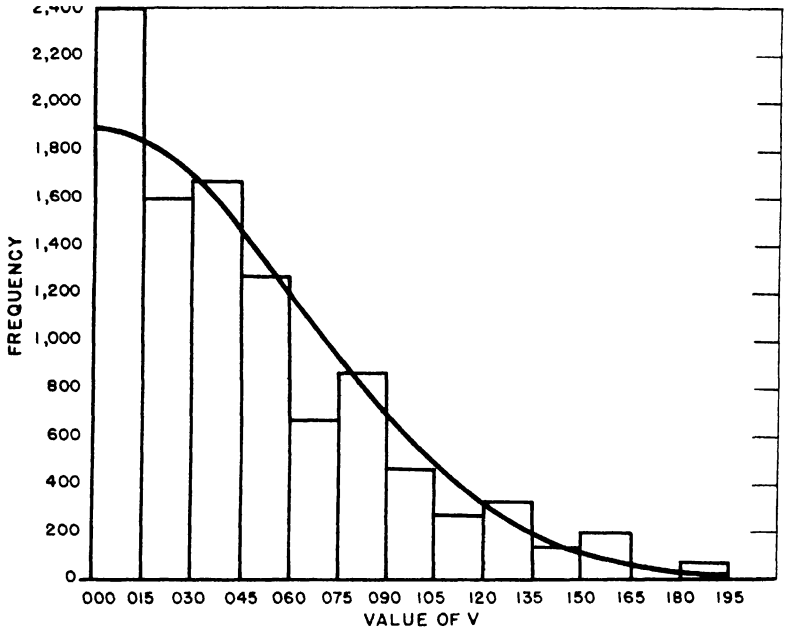


FIG. 2. OBSERVED AND THEORETICAL DISTRIBUTIONS OF VALUES OF v FOR 149 SAMPLES OF "DAYS TO MATURITY" IN RHODE ISLAND RED PULLETS FOR SAMPLES OF TWO OBSERVATIONS

mences, for samples of two observations made upon Rhode Island Red pullets. Figure 2 shows the observed distribution of the 149 coefficients of variation, together with the theoretical distribution given by equation (9).

In applying equation (9) to these data, the parameter, V, had to be evaluated from the data. The best estimates of the values of m, σ, and V which could be obtained from the 298 measurements of "days to maturity" are $m = 210.477$, $\sigma = 18.6991$; $V = 0.0888415$. The theoretical distribution shown in Figure 2 is based on this value of V.

The agreement between theory and observation shown by Figure 2 is very good. In this case, the chi test showed that the degree of discrepancy encountered could have arisen by chance about six times in ten trials.

SOME NOTES ON EXPONENTIAL ANALYSIS

By H. R. Grummann

Assistant Professor, Department of Applied Mathematics, Washington University

M. E. J. Geuhry de Bray in his charming little book "Exponentials made Easy"[1] tells how to determine the constants in the equation,

$$y = A_1\epsilon^{a_1 x} + A_2\epsilon^{a_2 x} \tag{I}$$

so that the curve will pass through four points, with equidistant ordinates on an empirical curve. If (Fig. 1) y_0, y_1, y_2, and y_3 are the equidistant ordinates and δ is their common separation, y_0 being the y intercept of the curve, de Bray's formulas are:

$$a_1 = \frac{\log z_1}{\delta}, \qquad a_2 = \frac{\log z_2}{\delta} \tag{II}$$

where z_1 and z_2 are the roots of the quadratic equation

$$\begin{vmatrix} z^2 & z & 1 \\ y_3 & y_2 & y_1 \\ y_2 & y_1 & y_0 \end{vmatrix} = 0. \tag{III}$$

The coefficients A_1 and A_2 of the two exponential terms are obtained by solving the two simultaneous equations

$$\begin{aligned} A_1 + A_2 &= y_0 \\ A_1 z_1 + A_2 z_2 &= y_1 \end{aligned} \tag{IV}$$

In attempting to find suitable empirical equations for some "river rating curves"—graphs of discharge versus stage—the writer tried to make use of de Bray's procedure. The original intention was to use the above method to determine the constants, and then to correct these constants by the use of Least Squares, as done by J. W. T. Walsh[2] in an application of the method to a problem in radioactivity. It often happens that a series of plotted observations suggest a simple exponential function, but that when the observations are replotted on semi-logarithmic paper a straight line is not obtained. Often, as in the case of a good many river rating curves, the result may be described

[1] Macmillan & Co. Ltd., St. Martin's St., London W. C. 2.

[2] Proceedings Phys. Soc. London XXXII. This reference is given by de Bray in his book, "Exponentials made Easy."

as "almost straight." At first blush it might seem that in all such cases it ought to be possible to fit a curve with equation I to the data by de Bray's Method. By an easy generalization of the above formulas, the constants in an equation with three or four exponential terms could be determined if two terms were not enough to secure a good fit.

It was soon found, however, that innocent looking monotonic curves without points of inflection plotted from data that gave an "almost straight" line on semi-logarithmic paper quite often led to a quadratic equation, (equation III) whose roots were not both positive numbers.

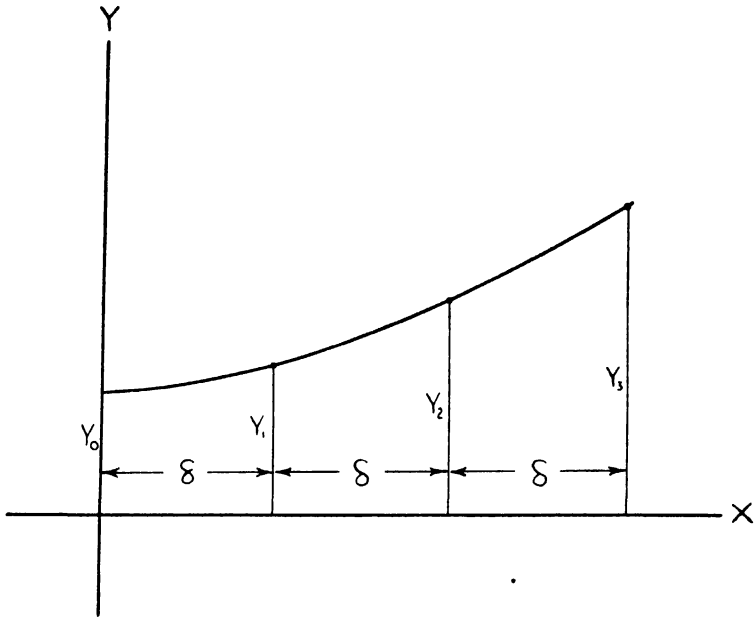


FIG. 1

If z_1 and z_2, the roots of III, are complex conjugates, it may be seen from IV that A_1 and A_2 will be complex conjugates. Also, a_1 and a_2 will be conjugate complex numbers and may be calculated as follows:

Let $z_1 = r\epsilon^{i\theta}$ and $z_2 = r\epsilon^{-i\theta}$
then from equation II,

$$r\epsilon^{i\theta} = \epsilon^{a_1\delta},$$

$$r\epsilon^{-i\theta} = \epsilon^{a_2\delta}$$

whence, by division to eliminate r we have

$$\epsilon^{2i\theta} = \epsilon^{\delta(a_1-a_2)}, \text{ or}$$

(Va)
$$\frac{2i\theta}{\delta} = a_1 - a_2.$$

Also, by multiplication to eliminate θ,

$$r^2 = \epsilon^{\delta(a_1+a_2)}, \text{ or}$$

$$\text{(Vb)} \qquad \frac{2 \log r}{\delta} = a_1 + a_2.$$

The sum and difference of the two a's being obtained by these expressions, one may solve for a_1 and a_2.

Let $\quad a_1 = \lambda + \iota\mu \qquad A_1 = \alpha + \iota\beta$

$$a_2 = \lambda - \iota\mu \qquad A_2 = \alpha - \iota\beta$$

Then equation I becomes

$$y = (\alpha + \iota\beta)\epsilon^{(\lambda+\iota\mu)x} + (\alpha - \iota\beta)\epsilon^{(\lambda-\iota\mu)x},$$

$$y = 2\epsilon^{\lambda x}[\alpha \cos \mu x - \beta \sin \mu x], \text{ or}$$

$$\text{(VI)} \qquad y = 2\epsilon^{\lambda x} R \cos (\mu x + c)$$

where $R = \sqrt{\alpha^2 + \beta^2}$ and $\tan c = \dfrac{\beta}{\alpha}$.

If one of the roots of III is negative, the de Bray formulas II and IV will still give an expression for equation I which formally reproduces y_0, y_1, y_2, and y_3 when 0, δ, 2δ, and 3δ, are substituted for x respectively, but which is useless for interpolating and of no value as a solution of the curve fitting problem. Suppose, for example, that z_1 is positive and z_2 is negative. Then

$$z_2 = (-1) \mid z_2 \mid \qquad \text{and}$$

$$\log z_2 = \log (-1) + \log \mid z_2 \mid.$$

Equation I then becomes

$$y = A_1\epsilon^{a_1 x} + (-1)^{\frac{x}{\delta}} A_2 \epsilon^{\frac{x \log |z_2|}{\delta}},$$

the factor $(-1)^{\frac{x}{\delta}}$ being real only when x is an integral multiple of δ. If the (-1) is written $e^{\iota\pi}$, we have

$$y = A_1\epsilon^{a_1 x} + \epsilon^{\frac{\pi \iota x}{\delta}} A_2 \epsilon^{\frac{x \log |z_2|}{\delta}}, \text{ or}$$

$$y = A_1\epsilon^{a_1 x} + A_2 \epsilon^{\frac{x \log |z_2|}{\delta}} \left[\cos \frac{\pi x}{\delta} + \iota \sin \frac{\pi x}{\delta}\right].$$

Neither the real nor the imaginary part would be a graduation function for a monotonic curve as each has a half period of δ.

The expression for I is similar, and of no greater practical value, if both of the roots of III are negative.

Without loss of generality we may let $y_0 = 1$, $r_1 = \frac{y_1}{y_0}$ $r_2 = \frac{y_2}{y_1}$ $r_3 = \frac{y_3}{y_2}$. Then the quadratic III becomes

$$\begin{vmatrix} z^2 & z & 1 \\ r_2 r_3 & r_2 & 1 \\ r_1 r_2 & r_1 & 1 \end{vmatrix} = 0, \quad \text{or,}$$

written in the form

$$z^2 + pz + q = 0, \quad \text{i.e.,}$$

$$\text{(IIIa)} \qquad z^2 + \frac{r_2(r_1 - r_3)}{(r_2 - r_1)} z + \frac{r_1 r_2 (r_3 - r_2)}{(r_2 - r_1)} = 0.$$

Hence the roots of this quadratic are real and unequal if $D > 6$, equal if $D = 6$, and complex if $D < 6$, where

$$D = \left[\frac{r_3}{r_1} - 3\frac{r_1}{r_3}\right] + 4\left[\frac{r_1}{r_2} + \frac{r_2}{r_3}\right]$$

From the point of view of the computer, however, it is about as much work to calculate D as to solve the quadratic equation.

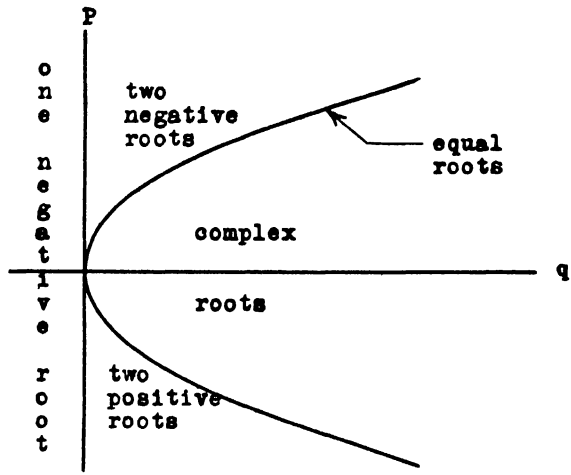


FIG. 2

Reverting to equation IIIa; suppose the numbers q and p are plotted as the coordinates of a point (q, p) as in Fig. 2. Then the parabola $p^2 = 4q$ is, so to speak, a locus of equal roots. The remainder of the figure requires no explanation.

Suppose that all the r's are positive, as they would be in the case of a simple monotonic curve which one proposed subjecting to an exponential analysis.

If $q < 0$, the quadratic will have one negative root. Now

$$q = \frac{r_1 r_2(r_3 - r_2)}{(r_2 - r_1)} \quad \text{and hence}$$

for $q < 0$, if $r_2 > r_1$, then $r_3 < r_2$ and consequently $r_3 < r_2 > r_1$ and if $r_2 < r_1$, then $r_3 > r_2$, or $r_1 > r_2 < r_3$. Also, provided $p_2 > 4q$, a positive p and a positive q will give two negative roots. But

$$p = \frac{r_2(r_1 - r_3)}{(r_2 - r_1)}$$

and p and q can not both be positive when all the r's are positive as this implies either that $r_2 > r_1$, $r_1 > r_3$ and $r_3 > r_2$, a contradiction, or else that $r_2 < r_1$, $r_1 < r_3$ and $r_3 < r_2$, also a contradiction. Hence if both roots are negative, the r's can not be all positive. The case of two negative roots will not arise in trying to fit equation I to a monotonic curve, since if all the r's are positive both p and q can not be positive.

For all r's positive, provided $p^2 > 4q$, a positive q and a negative p will give two positive roots. But

$$q = \frac{r_1 r_2(r_3 - r_2)}{(r_2 - r_1)} > 0,$$

and

$$-p = \frac{r_2(r_3 - r_1)}{(r_2 - r_1)} > 0$$

means that $r_3 > r_2 > r_1$ or $r_3 < r_2 < r_1$.

To sum up: If all the r's are positive, de Bray's method of exponential analysis is possible (a) when $D < 6$ and the roots of III are complex; (b) when $D > 6$ and $r_1 > r_2 > r_3$ or when $r_1 < r_2 < r_3$.

Figure 3 gives a picture of the second condition (b) of the preceding paragraph. Suppose an exponential curve is passed through the first *two* points on the empirical curve with ordinates y_0 and y_1. Its equation will be:

$$y = y_0 \left(\frac{y_1}{y_0}\right)^{\frac{x}{\delta}} = y_0 r_1^{\frac{x}{\delta}}.$$

Suppose also that y_2 is less than the ordinate to this curve when $x = 2\delta$. Now pass an exponential curve through y_1 and y_2 using a new axis of ordinates coinciding with y_1. Its equation is

$$y = y_1 \left(\frac{y_2}{y_1}\right)^{\frac{x}{\delta}} = y_1 r_2^{\frac{x}{\delta}},$$

or referred to the original axis:

$$y = y_1 r_2^{\frac{x-\delta}{\delta}}$$

Now if the graduation is possible without using trigonometric functions, y_3 must be less than the ordinate of this second curve when $x = 3\delta$.

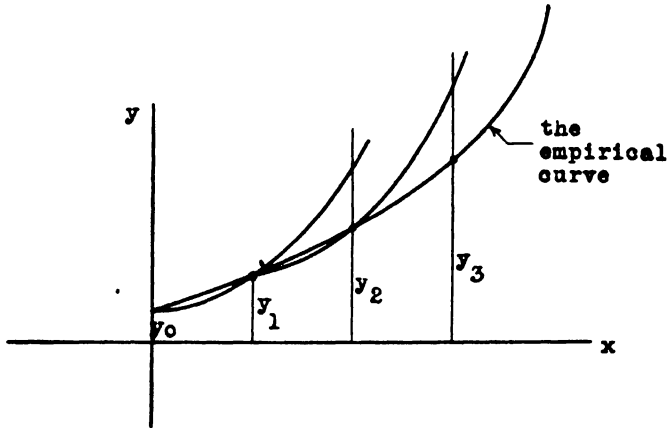


FIG. 3

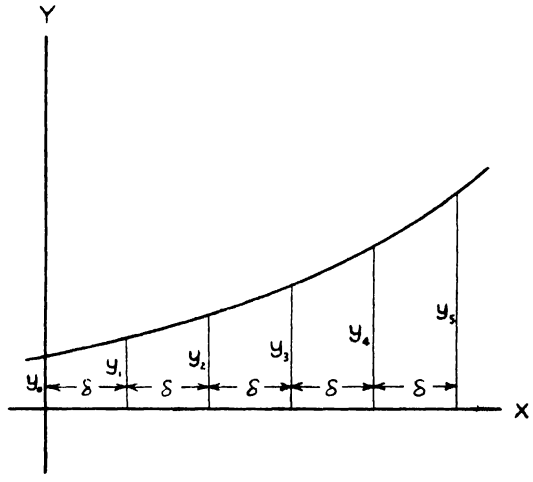


FIG. 4

It is natural to inquire if the state of affairs is not similar to this, for the cases of fitting curves with equations similar to I but having three or four exponential terms on the right hand side instead of only two. If three terms are used (see Fig. 4) to find constants in

(Ia) $$y = A_1 \epsilon^{\alpha_1 x} + A_2 \epsilon^{\alpha_2 x} + A_3 \epsilon^{\alpha_3 x}$$

it is first necessary to find the roots of the cubic

$$\text{(IIIa)} \qquad f(x) = \begin{vmatrix} z^3 & z^2 & z & 1 \\ y_5 & y_4 & y_3 & y_2 \\ y_4 & y_3 & y_2 & y_1 \\ y_3 & y_2 & y_1 & y_0 \end{vmatrix}.$$

Now, $f(x)$ will have no negative roots if $f(-x)$ has no changes of sign. But writing the conditions that the cofactors of the elements of the first row in the above determinant have the same signs, and assuming that all the y's are positive, one does not get a series of conditions analogous to $r_3 > r_2 > r_1$ or $r_3 < r_2 < r_1$.

In the following, formulas will be derived for finding the constants in equation Ia after the roots of IIIa have been determined. Also formulas will be obtained for finding the constants in

$$\text{(Ib)} \qquad y = A_1 \epsilon^{a_1 x} + A_2 \epsilon^{a_2 x} + A_3 \epsilon^{a_3 x} + A_4 \epsilon^{a_4 x}$$

after the roots of

$$\text{(IIIb)} \qquad \begin{vmatrix} z^4 & z^3 & z^2 & z & 1 \\ y_7 & y_6 & y_5 & y_4 & y_3 \\ y_6 & y_5 & y_4 & y_3 & y_2 \\ y_5 & y_4 & y_3 & y_2 & y_1 \\ y_4 & y_3 & y_2 & y_1 & y_0 \end{vmatrix} = 0$$

have been found. Both sets of formulas have been tested by an "exponential analysis" of the same body of data, viz., the very accurate recent determinations by the U. S. Bureau of Standards of the saturation pressure of water vapor above 100C.[3]

For the case of three exponential terms in the graduation function, the a's are found by formulas like II or V, after the roots of the cubic are found. If z_1, z_2, z_3 are the roots, the A's are obtained by solving the simultaneous equations

$$\text{(IVa)} \qquad \begin{aligned} A_1 + A_2 + A_3 &= y_0 \\ A_1 z_1 + A_2 z_2 + A_3 z_3 &= y_1 \\ A_1 z_1^2 + A_2 z_2^2 + A_3 z_3^2 &= y_2 \end{aligned}$$

[3] Osborne, Stimson, Fiock, and Ginnings: The Pressure of Saturated Water Vapor in the Range 100° to 374°C. Bureau Standards Journal of Research, Vol. 10, Febr. 1933, page 178.

This presents no new difficulty unless two of the roots are conjugate complex numbers. In this event, if we let z_1 = the real positive root, $z_2 = r\,\epsilon^{i\theta}$, and $z_3 = r\,\epsilon^{-i\theta}$ the determinant D of the equations IVa may be written

$$D = \begin{vmatrix} 1 & 1 & 1 \\ z_1 & r\epsilon^{i\theta} & r\epsilon^{-i\theta} \\ z_1^2 & r^2\epsilon^{2i\theta} & r^2\epsilon^{-2i\theta} \end{vmatrix}$$

or, expanded in terms of the elements of the first column and their minors,

$$D = 2i[z_1 r^2 \sin 2\theta - (r^3 + z_1^2 r)\sin\theta],$$

a pure imaginary. Similarly,

$$A_1 D = 2i[r^2 y_1 \sin 2\theta - (y_0 r^3 + y_2 r)\sin\theta],$$

also a pure imaginary, so that A_1 is real. Having calculated A_1, it is substituted in the first two of equations IVa, which are then solved for A_2 and A_3. a_2 and a_3 are then determined by formulas Va and Vb, replacing the subscripts 1 and 2 in those formulas, by the subscripts 2 and 3 respectively. Finally the two exponential terms corresponding to the complex roots of the cubic are combined into a single trigonometric term as in equation VI.

The necessary formulas for the case of four exponential terms in the graduation function will be discussed briefly. The equations

$$\text{(IVb)} \qquad \begin{aligned} A_1 + A_2 + A_3 + A_4 &= y_0 \\ A_1 z_1 + A_2 z_2 + A_3 z_3 + A_4 z_4 &= y_1 \\ A_1 z_1^2 + A_2 z_2^2 + A_3 z_3^2 + A_4 z_4^2 &= y_2 \\ A_1 z_1^3 + A_2 z_2^3 + A_3 z_3^3 + A_4 z_4^3 &= y_3 \end{aligned}$$

have to be solved for the A's. The z's are the roots of IIIb. Two cases will be considered: First case: z_1 and z_2 are complex conjugates and z_3 and z_4 are complex conjugates. Second case: z_1 and z_2 are complex conjugates and z_3 and z_4 are real and positive. In either event A_1 and A_2 are complex conjugates, as will be proved below. Formulas for A_1 are given for both cases. Then A_2 is known since it is the conjugate of A_1. Having found A_1 and A_2, let

$$c_0 = y_0 - (A_1 + A_2)$$

$$c_1 = y_1 - (A_1 z_1 + A_2 z_2)$$

Both c_0 and c_1 are then real. To get A_3 and A_4 solve the equations:

$$A_3 + A_4 = c_0$$

$$A_3 z_3 + A_4 z_4 = c_1$$

A pair of exponential terms with conjugate complex coefficients will then be expressed as a single real trigonometric term as in VI.

The determinant of equations IVb may be written

(VII) $$D = (z_1 - z_2)(z_1 - z_3)(z_1 - z_4)(z_2 - z_3)(z_2 - z_4)(z_3 - z_4).$$

First case: Let $z_1 = a + \iota b$, $z_2 = a - \iota b$, $z_3 = \alpha + \iota\beta$, $z_4 = \alpha - \iota\beta$. Then D may be written

(VIIa) $$D = -4\beta b[(a - \alpha)^2 + (b - \beta)^2]\,[(a - \alpha)^2 + (b + \beta)^2],$$

which is real. Now

$$A_1D + A_2D = \begin{vmatrix} y_0 & 1 & 1 & 1 \\ y_1 & z_2 & z_3 & z_4 \\ y_2 & z_2^2 & z_3^2 & z_4^2 \\ y_3 & z_2^3 & z_3^3 & z_4^3 \end{vmatrix} + \begin{vmatrix} 1 & y_0 & 1 & 1 \\ z_1 & y_1 & z_3 & z_4 \\ z_1^2 & y_2 & z_3^2 & z_4^2 \\ z_1^3 & y_3 & z_3^3 & z_4^3 \end{vmatrix}$$

$$= (z_1 - z_2) \begin{vmatrix} 0 & y_0 & 1 & 1 \\ 1 & y_1 & z_3 & z_4 \\ (z_1 + z_2) & y_2 & z_3^2 & z_4^2 \\ (z_1^2 + z_1z_2 + z_2^2) & y_3 & z_3^3 & z_4^3 \end{vmatrix}$$

and this is real since $(z_1 - z_2)$ is a pure imaginary and the minors of the real elements of the first column of the determinant are all pure imaginaries. Hence A_1 and A_2 are complex conjugates since when each is expressed as a quotient of two determinants by Cramer's rule, the sum of the two numerators is real and the common denominator is also real.

For purposes of numerical calculation A_1 may be obtained from

$$A_1 = \frac{NP}{D}$$

in which D is obtained from VIIa,

$$N = y_3 - (z_2 + z_3 + z_4)y_2 + (z_2z_3 + z_2z_4 + z_3z_4)y_1 - (z_2z_3z_4)y_0,$$

$$\text{and } P = (z_2 - z_3)(z_2 - z_4)(z_3 - z_4)$$

$$= 2\beta[(\alpha - a)2b + \iota\{(\alpha - a)^2 + (\beta^2 - b^2)\}], \text{ a complex number.}$$

If $z_1z_2 = r^2$ and $z_3z_4 = \rho^2$, the symmetric functions of the z's in the above formula may be calculated from

$$z_2z_3z_4 = (a - \iota b)\rho^2$$

$$z_2z_3 + z_2z_4 + z_3z_4 = \rho^2 + 2\alpha(a - \iota b)$$

$$z_2 + z_3 + z_4 = (a - \iota b) + 2\alpha$$

For the second case, which is exemplified by the vapor pressure data,

$$D = 2\iota b[(a - z_3)^2 + b^2]\,[(a - z_4)^2 + b^2]\,[z_3 - z_4], \tag{VIIb}$$

a pure imaginary. The sum of the two numerators of A_1 and A_2, namely

$$(z_1 - z_2)\begin{vmatrix} 0 & y_0 & 1 & 1 \\ 1 & y_1 & z_3 & z_4 \\ z_1 + z_2 & y_2 & z_3^2 & z_4^2 \\ z_1^2 + z_1z_2 + z_2^2 & y_3 & z_3^3 & z_4^3 \end{vmatrix}$$

is a pure imaginary, since $(z_1 - z_2)$ has this character, and the determinant has nothing but real elements. Hence A_1 and A_2 are still complex conjugates when z_3 and z_4 are real, z_1 and z_2 being complex conjugates.

For purposes of numerical calculation A_1 may be obtained from

$$A_1 = \frac{N}{(z_1 - z_2)\,(z_1 - z_3)\,(z_1 - z_4)}.$$

Here $(z_1 - z_2)$ is a pure imaginary and the other three factors are complex.

Let

$$N = r_1(\cos\theta_1 + \iota\sin\theta_1)$$

$$z_1 - z_3 = r_2(\cos\theta_2 + \iota\sin\theta_2)$$

$$z_1 - z_4 = r_3(\cos\theta_3 + \iota\sin\theta_3)$$

Then

$$A_1 = \frac{r_1\,[\cos(\theta_1 - \theta_2 - \theta_3) + \iota\sin(\theta_1 - \theta_2 - \theta_3)]}{(z_1 - z_2)\,r_2\,r_3}$$

In calculating N by the formula given for it in the preceding paragraph, the symmetric functions of the z's were obtained from

$$z_2z_3z_4 = (a - \iota b)z_3z_4$$

$$z_2z_3 + z_2z_4 + z_3z_4 = (a - \iota b)(z_3 + z_4) + z_3z_4$$

$$z_2 + z_3 + z_4 = (a - \iota b) + z_3 + z_4.$$

Example

The first two of the following tables are abstracted from Table 2, p. 178 of Bureau Standards Research Paper No. 523. The third table is abstracted from Table 3, p. 179 et. seq. of that publication. x is the number of degrees centigrade *above* 100°. y is the pressure of saturated water vapor in International Standard Atmospheres. In the first two of the following tables, the values

of y are observed values. In the third, they are interpolated or graduated values calculated at the Bureau of Standards.

TABLE I

x	y
0	1.0000
90	12.3887
180	63.3558
270	207.771

TABLE II

x	y
0	1.0000
50	4.6969
100	15.3472
150	39.2566
200	84.7969
250	163.205

TABLE III

x	y
0	1.0000
39	3.4666
78	9.4490
117	21.612
156	43.392
195	78.974
234	133.64
273	215.37

The observed values of y in Table I are reproduced by the following formula used in conjunction with a standard six place table of logarithms and trigonometric functions:

(I) $$y = 3.967433\ \epsilon^{.01539540x} \cos (.4085758x - 75°24'03''.7).$$

The observed values of y in Table II are reproduced by the following formula used in conjunction with a standard six place table of logarithms and trigonometric functions.

(II) $$y = 3.0253744\ \epsilon^{.01515605x} + 2.2171657\ \epsilon^{.0115007 16x} \cos (155°59'35''.5 - 0.7899232x).$$

Hence the formula is presumably an excellent one for interpolation between the values of y listed in Table II, if the greatest accuracy is not needed.[4]

The values of y in Table III are reproduced exactly to five significant figures by the following formula used in conjunction with a standard six place table of logarithms and trigonometric functions.

$$y = 3.8902543\,\epsilon^{.01413920x} - .164787\,\epsilon^{-.0216980x}$$

$$+ 2.743000\,\epsilon^{.009884290x}\cos(.7860725x + 186°28'53''.2).$$

By means of this formula the saturation pressure of water vapor was calculated for every five degrees from 100°C to 370°C in order to make comparisons with the corresponding "smoothed" values in Table 2 of the Bureau of Standards publication referred to above. The discrepancies were never more than one in the fourth significant figure and generally less. The poorest agreement was in the ranges of temperature from 100°C to 135°C and from 245°C to 270°C.

It is a pleasure to acknowledge the intelligent and painstaking assistance of Mr. G. D. Lambert, undergraduate student at Washington University, for doing most of the computing.

WASHINGTON UNIVERSITY,
ST. LOUIS, MO.

[4] The values of y in Table III (not counting the value of y for $x = 0$) are reproduced by it with an average error of .13% and a largest error (for $x = 234°$) of .30%. Four of the errors are negative and three positive.

ON THE FREQUENCY DISTRIBUTION OF CERTAIN RATIOS

By H. L. Rietz
University of Iowa

Considerable interest in the distribution of ratios, $t = y/x$, has no doubt been suggested by important applications. For example, we may mention the opsonic index in bacteriology, the ratio of systolic to diastolic blood pressure in physiology, and ratios such as link relatives or certain index numbers in economics.

In 1910, Karl Pearson[1] gave certain properties of the distribution of ratios by means of approximate formulas for moments up to order four in terms of means, variances, product moments, and coefficients of variability of x and y. The resulting formulas did not give, with sufficient accuracy, the constants of the distribution of the opsonic index for the purpose of Dr. Greenwood to whom Pearson attributed the derivation of the formulas for the special case in which x and y are uncorrelated. Pearson next adopted the plan of tabulating the reciprocals, say $x' = \frac{1}{x}$, and then finding the constants of the distribution of the product yx' in the case in which x' and y are uncorrelated. He then obtained satisfactory results in illustrative examples.

In 1929, C. C. Craig[2] obtained the semi-invariants of y/x in terms of moments of x and y, and then expressed the moments in terms of the semi-invariants of the distribution function, $f(x, y)$, of x and y. By this means, he was able to deal with the case in which x and y are normally correlated under suitable conditions. Craig found it desirable to restrict the distribution of x in such a way that the probability of a zero value of x is an infinitesimal of sufficiently high order that a certain integral exists. This limitation seems to imply in applications to actual data that no zero values of x are to occur. This suggests that we deal with the cases of x at or near zero with considerable care.

By starting with the assumption that the values of x and y are a set of normally distributed pairs of values with correlation coefficient r, and by considering the quotient $z = \frac{b + y}{a + x}$, a and b being constants, R. C. Geary,[3] in a paper published in 1930, found an algebraic function, $u = f(z)$, of fairly simple form with the property that u is nearly normally distributed with arithmetic mean zero and standard deviation unity provided that $a + x$ is unlikely to

[1] On the constants of index distributions, Biometrika, Vol. 7 (1910), pp. 531–546.

[2] The frequency function of y/x, Annals of Mathematics, Vol. 30 (1928–29), pp. 471–486.

[3] The frequency distribution of the quotient of two normal variates, J. Royal Statistical Society, Vol. XCIII (1930), pp. 442–7.

have negative values. Here we have again a suggestion to exercise special care in the case of quotients with the divisor near zero or negative.

In 1932, Fieller[4] obtained in explicit form the approximate distribution of $t = y/x$ where values (x, y) are drawn from the bivariate normal distribution

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}\, e^{-\frac{1}{2}\frac{1}{1-r^2}\left\{\frac{(x-\bar{x})^2}{\sigma_x^2}+\frac{(y-\bar{y})^2}{\sigma_y^2}-2r\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right\}}$$

under the condition that $\bar{x}$ is large compared with σ_x.

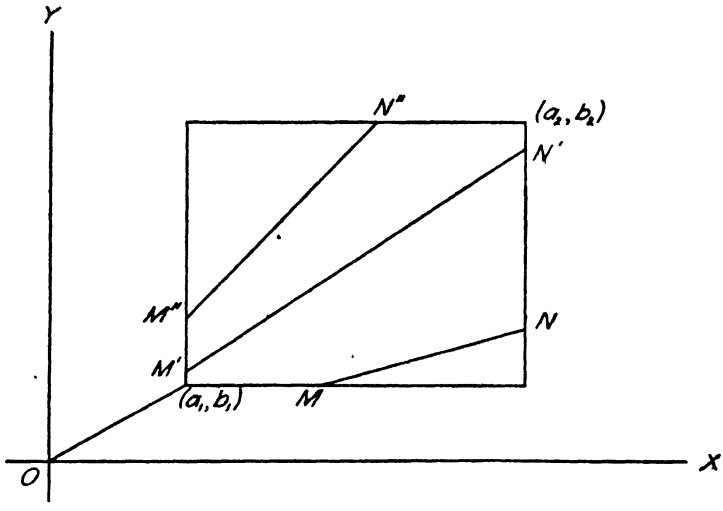


FIG. 1

Very recently Kullback[5] found the distribution law of the quotient, $t = y/x$, where x and y are drawn from Pearson Type III parent populations given by

$$f_1(x) = \frac{e^{-x}x^{p-1}}{\Gamma(p)}; \qquad f_2(y) = \frac{e^{-y}y^{q-1}}{\Gamma(q)}, \qquad 0 \leqq x \leqq \infty, \quad 0 \leqq y \leqq \infty.$$

It is fairly easy to see, in a general way, that the distribution of $t = y/x$ depends very much on the location of the origin as well as on the parent distribution from which x and y are drawn. This fact will be fairly obvious from the present paper whose main purpose is to give clear geometrical descriptions of the distributions of ratios, $t = y/x$, for each of several cases in which (x, y) are points taken at random from certain simple geometrical figures conveniently located with respect to the origin.

In accord with the suggestions to be cautious when the divisor is near zero or negative, we consider first the very simple case of ratios $t = y/x$ obtained

[4] E. C. Fieller, The distribution of the index in a normal bivariate population, Biometrika, Vol. 24 (1932), pp. 428–440.

[5] Solomon Kullback, Annals of Mathematical Statistics, Vol. VII (1936), pp. 51–53.

from points uniformly distributed over a rectangle such as is shown in Fig. 1 with sides parallel to coordinate axes and $a_1 > 0$, $b_1 > 0$. As indicated on Fig. 1, we assume for simplicity that the coordinates of the points are positive and $a_1 \leqq x \leqq a_2$, $b_1 \leqq y \leqq b_2$.

Case I. When $\frac{b_1}{a_1} \leqq \frac{b_2}{a_2}$, Fig. 1.

Let $k\,dx\,dy$ be the probability that a point (x, y) taken at random in the rectangle will fall into $dxdy$ where k is a constant. Then

$$k \int_{b_1}^{b_2} \int_{a_1}^{a_2} dxdy = k(a_2 - a_1)(b_2 - b_1) = 1,$$

and

$$k = \frac{1}{(a_2 - a_1)(b_2 - b_1)}.$$

Transform the element $k\,dxdy$ into one with variables t, and x by making

$$x = x,$$

$$y = tx.$$

The Jacobian is $|x| = x$.

The new element is $k\,x\,dxdt$ and is to be integrated over the range on x for an assigned t in order to get the probability, to within infinitesimals of higher order, that a random t falls into an assigned dt. By assigning t any value such that $\frac{b_1}{a_2} \leqq t \leqq \frac{b_1}{a_1}$, say t is the slope of MN, (Fig. 1), we have

$$(1) \qquad k \int_{\frac{b_1}{t}}^{a_2} x dx dt = \frac{k}{2}\left(a_2^2 - \frac{b_1^2}{t^2}\right) dt$$

the limits of integration being indicated by the ends of the line MN.

When the assigned t is such that $\frac{b_1}{a_1} \leqq t \leqq \frac{b_2}{a_2}$, say t is the slope of the line $M'N'$, we have

$$(2) \qquad k \int_{a_1}^{a_2} x\,dx\,dt = \frac{k}{2}(a_2^2 - a_1^2)\,dt$$

When the assigned t is such that $\frac{b_2}{a_2} \leqq t \leqq \frac{b_2}{a_1}$, say it is the slope of $M''N''$, we have

$$(3) \qquad k \int_{a_1}^{\frac{b_2}{t}} x\,dx\,dt = \frac{k}{2}\left(\frac{b_2^2}{t^2} - a_1^2\right) dt$$

Thus, from (1), (2), (3), when as in Fig. 1, $\frac{b_1}{a_1} \leqq \frac{b_2}{a_2}$, the frequency function of t is given by

$$F(t) = \frac{k}{2}\left(a_2^2 - \frac{b_1^2}{t^2}\right) \quad \text{when} \quad \frac{b_1}{a_2} \leqq t \leqq \frac{b_1}{a_1}, \tag{4}$$

$$F(t) = \frac{k}{2}(a_2^2 - a_1^2) \quad \text{when} \quad \frac{b_1}{a_1} \leqq t \leqq \frac{b_2}{a_2}, \tag{5}$$

$$F(t) = \frac{k}{2}\left(\frac{b_2^2}{t^2} - a_1^2\right) \quad \text{when} \quad \frac{b_2}{a_2} \leqq t \leqq \frac{b_2}{a_1}. \tag{6}$$

See Fig. 2 for the general form of the frequency curve $F(t)$ when $\frac{b_1}{a_1} < \frac{b_2}{a_2}$ with the segment from $t = \frac{b_1}{a_1}$ to $\frac{b_2}{a_2}$ a horizontal straight line and with discontinuities in the first derivatives of $F(t)$ at $t = \frac{b_1}{a_1}$ and $t = \frac{b_2}{a_2}$.

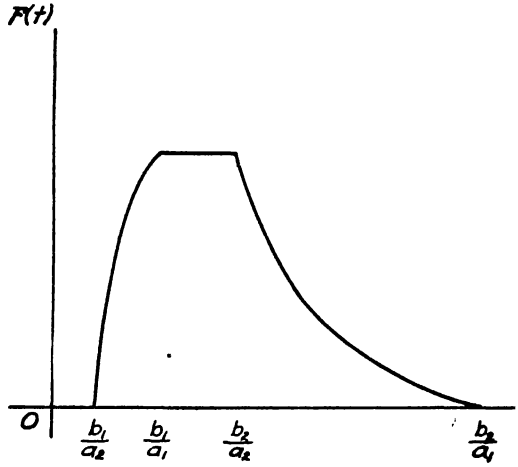


FIG. 2

When $a_1 \to 0$, and $b_1 = 0$, the frequency curve approaches

$$F(t) = \frac{a_2}{2b_2} \quad \text{when} \quad 0 \leqq t \leqq \frac{b_2}{a_2} \tag{7}$$

$$F(t) = \frac{b_2}{2a_2t^2} \quad \text{when} \quad t \geqq \frac{b_2}{a_2}. \tag{8}$$

It may be noted that the curve given by making $a_1 = 0$ and $b_1 = 0$ extends to infinity, and that the first and second moments about the origin are each infinite.

Case II. When $\frac{b_1}{a_1} > \frac{b_2}{a_2}$.

If the rectangle in Fig. 1 were moved upward keeping its sides parallel to the x and y axes until $\frac{b_1}{a_1} > \frac{b_2}{a_2}$, we would obtain

$$F(t) = \frac{k}{2}\left(a_2^2 - \frac{b_1^2}{t^2}\right) \quad \text{if} \quad \frac{b_1}{a_2} \leqq t \leqq \frac{b_2}{a_2}, \tag{9}$$

$$F(t) = \frac{k}{2t^2}(b_2^2 - b_1^2) \quad \text{if} \quad \frac{b_2}{a_2} \leqq t \leqq \frac{b_1}{a_1}, \tag{10}$$

$$F(t) = \frac{k}{2}\left(\frac{b_2^2}{t^2} - a_1^2\right) \quad \text{if} \quad \frac{b_1}{a_1} \leqq t \leqq \frac{b_2}{a_1}. \tag{11}$$

By comparing (5) and (10), it may be observed that $F(t)$ of the middle segment of the distribution curve differs much in Case II from its corresponding constant value in Case I.

By moving the rectangle of Fig. 1 downward, keeping its sides parallel to the x and y axes until b_1 is negative, we easily find further forms of the distribution curve $F(t)$.

To consider the distribution of the ratio $t = y/x$ for another very simple type of distribution of x and y, suppose we have given the distribution function

$$f(x, y) = k\, e^{-\frac{x}{a} - \frac{y}{b}}, \quad \begin{pmatrix} x \geqq c > 0, \ y \text{ non-negative} \\ a > c, \ b > 0 \end{pmatrix} \tag{12}$$

where $\int_0^\infty \int_c^\infty f(x, y)\, dx\, dy = 1$. Then

$$k = \frac{e^{c/a}}{ab}.$$

In this case,

$$\begin{aligned} F(t) &= \frac{e^{c/a}}{ab} \int_c^\infty x\, e^{-\frac{x}{a} - \frac{xt}{b}}\, dx \\ &= \frac{1}{b + at}\left(c + \frac{ab}{b + at}\right) e^{-\frac{ct}{b}}, \end{aligned} \tag{13}$$

a monotone decreasing function from $t = 0$ to $t = \infty$.

With $c = 0$ as a limiting value, we obtain

$$F(t) = \frac{ab}{(b + at)^2}, \tag{14}$$

a distribution curve with the mean value of t at infinity.

If we should similarly consider

$$f(x, y) = \frac{2}{\pi\sigma_x\sigma_y} e^{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}} \quad (x \text{ and } y \text{ non-negative}) \tag{15}$$

we easily obtain

$$F(t) = \frac{1}{2\pi\sigma_x\sigma_y\left(\frac{1}{\sigma_x^2} + \frac{t^2}{\sigma_y^2}\right)} \tag{16}$$

as the distribution function.

Although the difficulties[6] of the problem of the distribution of the ratio y/x when x and y are normally correlated have been overcome[7] to a considerable

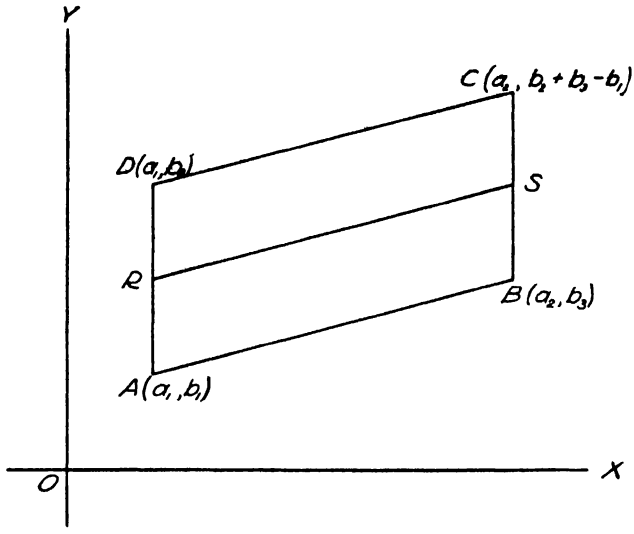


FIG. 3

extent, still the examination of some very simple geometric cases of non-normal but linear correlation may not be without some interest. Such a case will now be considered.

For one very simple case in which x and y are correlated, suppose we are given a set of points (x, y) uniformly distributed over the parallelogram $ABCD$ (Fig. 3) with sides AD and BC parallel to the y-axis so that the regression of y on x is linear as shown by the line RS.

The equation of RS is

$$y = m(x - a_1) + \frac{b_1 + b_2}{2}. \tag{17}$$

[6] Loc. cit., Pearson, p. 531.

[7] Loc. cit., C. C. Craig, R. C. Geary, E. C. Fieller.

Then although x_i and y_i are correlated, x_i and

$$y_i' = y_i - m(x_i - a_i) - \frac{b_1 + b_2}{2}$$

are uncorrelated. Let us consider the distribution of the ratio $t' = \frac{y_i'}{x_i}$.

Consider the element of frequency $kdxdy'$, where

$$k(b_2 - b_1)(a_2 - a_1) = 1. \tag{18}$$

Change variables to x and t' by the transformation

$$x = x,$$
$$y' = t'x.$$

Then the element of frequency becomes

$$kx\, dx\, dt'. \tag{19}$$

Next integrate (19) with respect to x under the restriction that t' is assigned. Three cases occur:

(a) When $-\frac{b_2 - b_1}{2a_2} \leqq t' \leqq \frac{b_2 - b_1}{2a_2}$, we obtain by integration of (19) for the element of relative frequency of t' in dt',

$$k\int_{a_1}^{a_2} x\, dx\, dt' = \frac{k}{2}(a_2^2 - a_1^2)\, dt'. \tag{20}$$

(b) When $t' \geqq \frac{b_2 - b_1}{2a_2}$, we obtain

$$k\int_{a_1}^{\frac{b_2-b_1}{2t'}} x\, dx\, dt' = \frac{k}{2}\left[\frac{(b_2 - b_1)^2}{4t'^2} - a_1^2\right] dt' \tag{21}$$

(c) When $t' \leqq -\frac{b_2 - b_1}{2a_2}$, we similarly obtain

$$k\int_{a_1}^{-\frac{b_2-b_1}{2t'}} x\, dx\, dt' = \frac{k}{2}\left[\frac{(b_2 - b_1)^2}{4t'^2} - a_1^2\right] dt' \tag{22}$$

From (18), (19), (20), (21) and (22), the frequency function of t' is given by

$$F(t') = \frac{a_2 + a_1}{2(b_2 - b_1)} \quad \text{when} \quad -\frac{b_2 - b_1}{2a_2} \leqq t' \leqq \frac{b_2 - b_1}{2a_2}; \tag{23}$$

$$F(t') = \frac{1}{2(b_2 - b_1)(a_2 - a_1)}\left[\frac{(b_2 - b_1)^2}{4t'^2} - a_1^2\right], \tag{24}$$

where the range of t' is subject to either the inequalities,

$$\frac{b_2 - b_1}{2a_2} \leqq t' \leqq \frac{b_2 - b_1}{2a_1}, \quad \text{or} \quad -\frac{b_2 - b_1}{2a_1} \leqq t' \leqq -\frac{b_2 - b_1}{2a_2}.$$

See Fig. 4 for the general form of the $F(t')$ frequency curve.
If we make $a_1 = 0$, the curve becomes infinite in range. If we make not only $a_1 = 0$, but $(b_1 + b_2)/2 = 0$, we have, in place of (17),

$$y = mx.$$

In this limiting situation, if we make $a_2 = a$ and $\frac{b_2 - b_1}{2} = b$,

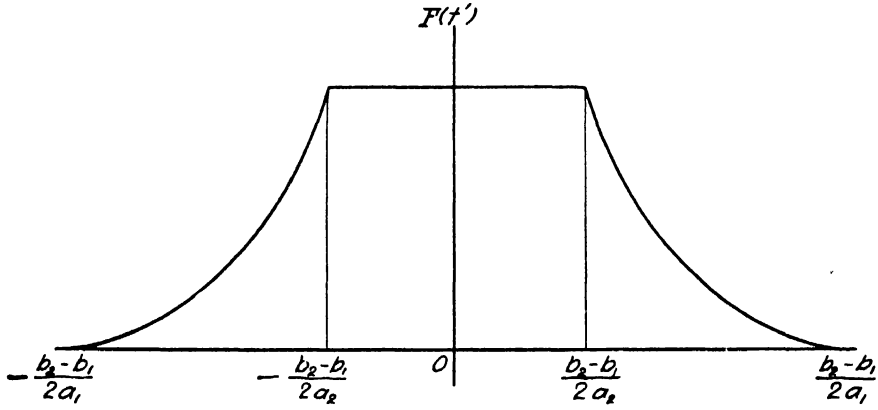


FIG. 4

(23) becomes

(25) $$F(t') = \frac{a}{4b^2}, \quad \text{for} \quad -\frac{b}{a} \leqq t' \leqq \frac{b}{a}, \text{ and (24) becomes}$$

(26) $$F(t') = \frac{b}{4\,at'^2} \quad \text{for} \quad t' \geqq \frac{b}{a} \quad \text{and for} \quad t' \leqq -\frac{b}{a}.$$

Then we have $y' = y - mx$

and $$t' = \frac{y'}{x} = t - m.$$

Further, if t' is distributed in accord with a frequency function, $F(t')$, the distribution of $t = t' + m$ with m constant is given by

$$F(t - m).$$

Hence, the probability that a random value t will fall into a range t to $t + dt$ is given to within infinitesimals of higher order by

$$\frac{a}{4b}\,dt \quad \text{when} \quad m - \frac{b}{a} \leqq t \leqq m + \frac{b}{a}, \tag{27}$$

and by

$$\frac{b\,dt}{4a(t-m)^2} \quad \text{when} \quad t \geqq m + \frac{b}{a} \text{ and } t \leqq m - \frac{b}{a}. \tag{28}$$

With the frequency curve given by (27) and (28) we may note that the variance of t becomes infinite.

Without taking the space to continue illustrations, it is fairly obvious that a wide diversity of form can be given to the frequency function of the quotients $t = y/x$ by relatively simple changes in the location of a sample parent population with reference to the origin.

EDITORIAL

THE FUNDAMENTAL NATURE AND PROOF OF SHEPPARD'S ADJUSTMENTS

In the course of our discussion of moment adjustments, we shall have occasion to refer to the following lengthy distribution of discrete variates. By selecting

TABLE 1

*Distribution of the number of items correctly recorded by 244 students in a five minute code transcription test**

Score x	Freq. f	Score x	Freq. f	Score x	Freq. f
64	1	94	3	119	1
66	2	95	5	120	2
68	2	96	3	121	6
69	1	97	3	122	2
70	1	98	12	123	3
71	3	99	4	124	2
72	3	100	5	125	6
73	3	101	6	126	3
76	1	102	8	127	4
77	2	103	6	128	2
78	3	104	8	130	2
79	1	105	9	131	1
80	2	106	5	132	5
82	2	107	3	133	1
83	3	108	3	134	1
84	2	109	4	136	1
85	6	110	2	138	1
86	3	111	4	140	1
87	1	112	7	141	1
88	2	113	5	142	2
89	4	114	5	144	2
90	4	115	7	153	1
91	5	116	8	155	1
92	2	117	3		
93	4	118	2	Total	244

* I am indebted to Professor J. A. Gengerelli, of the Department of Psychology of Univ. of California at Los Angeles, for these data.

the provisional mean, $M_0 = 105$, we find that

$$\Sigma x f = -129 \qquad \Sigma x^3 f = -52\ 005$$

$$\Sigma x^2 f = 77\ 591 \qquad \Sigma x^4 f = 69\ 239\ 951.$$

Let us now form the nine possible distributions of grouped-discrete variates that arise from the nine possible "groupings of nine." These are presented in table 2.

TABLE 2

Distributions derived from the data of table 1 by making the nine possible "groupings of nine"

First significant class interval of distribution								
(1) 64–72	(2) 63–71	(3) 62–70	(4) 61–69	(5) 60–68	(6) 59–67	(7) 58–66	(8) 57–65	(9) 56–64
13	10	7	6	5	3	3	1	1
12	15	16	16	14	14	13	15	15
27	23	21	20	22	21	16	14	11
41	41	33	32	30	28	31	29	30
53	54	63	61	55	52	49	45	41
45	45	40	38	42	45	44	48	52
27	27	29	34	36	39	40	42	43
16	19	24	25	23	24	28	30	29
8	6	7	6	10	10	12	11	13
1	2	2	4	5	6	6	7	7
1	2	2	2	2	2	2	2	1
								1

Let us now compute the values of Σxf, $\Sigma x^2 f$, $\Sigma x^3 f$ and $\Sigma x^4 f$ for each of the distributions of table 2, selecting $M_0 = 105$ in each instance in order to facilitate a comparison of these results with those for table 1. Thus, in spite of what would otherwise be called poor computing technique, we shall use the following class marks as values of x for the first distribution above; $-37, -28, -19, \cdots, 35, 44, 53$. For the second we shall likewise use, $-38, -29, -20, \cdots, 34, 43, 52$, respectively.

TABLE 3

Summations derived from the distributions listed in table 2, using $M_0 = 105$

Dist.	Σxf	$\Sigma x^2 f$	$\Sigma x^3 f$	$\Sigma x^4 f$
(1)	— 181	77 149	— 134 191	69 063 265
(2)	— 218	78 466	— 54 602	74 519 962
(3)	— 111	77 769	2 889	71 465 409

TABLE 3—*Continued*

Dist.	Σxf	Σx^2f	Σx^3f	Σx^4f
(4)	— 139	79 747	— 23 311	74 171 443
(5)	— 104	81 934	19 666	76 143 874
(6)	— 87	80 145	16 551	72 467 541
(7)	— 52	80 302	— 36 118	71 851 930
(8)	— 89	78 553	— 101 357	68 426 497
(9)	— 180	78 894	— 180 792	73 155 150
Average	— 129	79 217⅔	— 54 585	72 362 785⅔

The fact that the average of the values of Σxf appearing in table 3 suggests that no adjustments of the first moment is necessary and that the variations in the nine values for Σxf may be regarded as *accidental errors* and attributed to grouping. An attempt to account for this phenomenon and also for the fact that the averages of the higher order summations of table 3 do not likewise agree with the corresponding summations of table 1 lead us directly to formulae for Sheppard's adjustments.

For the moment, let us concentrate our intention upon a single variate, x_0, and its associated frequency, f_{x_0}, that are a part of a distribution of discrete variates, such as table 1. Suppose we were to form the k different distributions arising from the k possible "groupings of k." In one of these distributions, x_0 will rest in the first position of a class interval: the limits of this class are x_0 and $(x_0 + k - 1)$ and the class mark is therefore $[x_0 + \frac{1}{2}(k - 1)]$. The contribution of the variate, x_0, to $\Sigma x^s f$ for this particular distribution is therefore

$$[x_0 + \tfrac{1}{2}(k - 1)]^s \cdot f_{x_0}.$$

If x_0 rests in the second position of a class, the limits of this class will be $(x_0 - 1)$ and $(x_0 + k - 2)$ and the corresponding class mark is $[x_0 + \frac{1}{2}(k - 3)]$ and the contribution of x_0 to $\Sigma x^s f$ for this distribution is

$$[x_0 + \tfrac{1}{2}(k - 3)]^s \cdot f_{x_0}.$$

The *expected* value of $\Sigma x^s f$ arising from the k different groupings of variates is therefore,

$$(1) \qquad E(\textstyle\sum x^s f) = \frac{1}{k}\left[\sum^{1} x^s f + \sum^{2} x^s f + \cdots + \sum^{k} x^s f\right]$$

where $\sum^{i} x^s f$ refers to that distribution in which a specified x_0 rests in the i-th position in the class in which it occurs. The contribution of x_0 to this expected value is therefore

$$(2)\quad \frac{1}{k}\{[x_0 + \tfrac{1}{2}(k-1)]^s + [x_0 + \tfrac{1}{2}(k-3)]^s + [x_0 + \tfrac{1}{2}(k-5)]^s + \cdots\} f_{x_0},$$

this series consisting obviously of k terms.

Expanding each term of (2) by the binomial theorem yields

$$\frac{1}{k}\left[x_0^s - {}_sC_1\, x_0^{s-1}\left(\frac{k-1}{2}\right) + {}_sC_2\, x_0^{s-2}\left(\frac{k-1}{2}\right)^2 - {}_sC_3\, x_0^{s-3}\left(\frac{k-1}{2}\right)^3 + \cdots\right]$$

$$\frac{1}{k}\left[x_0^s - {}_sC_1\, x_0^{s-1}\left(\frac{k-3}{2}\right) + {}_sC_2\, x_0^{s-2}\left(\frac{k-3}{2}\right)^2 - {}_sC_3\, x_0^{s-3}\left(\frac{k-3}{2}\right)^3 + \cdots\right]$$

$$\frac{1}{k}\left[x_0^s - {}_sC_1\, x_0^{s-1}\left(\frac{k-5}{2}\right) + {}_sC_2\, x_0^{s-2}\left(\frac{k-5}{2}\right)^2 - {}_sC_3\, x_0^{s-3}\left(\frac{k-5}{2}\right)^3 + \cdots\right]$$

etc.

Since s is an integer, series (2) may be written as the sum of the $(s+1)$ terms of the series

$$(3)\qquad [x_0^s\, S_0 - {}_sC_1\, x_0^{s-1}\, S_1 + {}_sC_2\, x_0^{s-2}\, S_2 - {}_sC_3\, x_0^{s-3}\, S_3 + \cdots] f_{x_0},$$

where

$$S_i = \frac{1}{k}\left[\left(\frac{k-1}{2}\right)^i + \left(\frac{k-3}{2}\right)^i + \left(\frac{k-5}{2}\right)^i + \cdots \text{ to } k \text{ terms}\right].$$

By the Euler-Maclaurin Sum Formula we have

$$\sum_{x=a}^{b} x^p = \frac{1}{p+1}(b^{p+1} - a^{p+1}) + \tfrac{1}{2}(b^p + a^p) + \frac{B_1}{2!}\, p\,(b^{p-1} - a^{p-1})$$

$$- \frac{B_3}{4!}\, p^{(3)}\,(b^{p-3} - a^{p-3}) + \frac{B_5}{6!}\, p^{(5)}\,(b^{p-5} - a^{p-5}) + \cdots,$$

where $p^{(i)} = p(p-1)(p-2)(p-3)\cdots$ to i factors. In our expression for S_i, $a = \frac{1}{2}(k-1) = -b$, and therefore S_i equals zero when i is an odd integer. For even values of i,

$$(4)\qquad S_i = \frac{2}{k}\left\{\frac{(k-1)^i\,(k+i)}{2^{i+1}\,(1+i)} + \frac{B_1}{2!}\, i\left(\frac{k-1}{2}\right)^{i-1}\right.$$

$$\left. \frac{B_3}{4!}\, i^{(3)}\left(\frac{k-1}{2}\right)^{i-3} + \frac{B_5}{6!}\, i^{(5)}\left(\frac{k-1}{2}\right)^{i-5} - \cdots\right\}$$

so that

$$S_0 = 1$$

$$S_2 = \frac{1}{12}(k^2 - 1)$$

$$S_4 = \frac{1}{240}(k^2 - 1)(3k^2 - 7)$$

$$S_6 = \frac{1}{1344}(k^2 - 1)(3k^4 - 18k^2 + 31)$$

etc.

Since expression (3) represents the contribution of any variate, x_0, to the expected value defined by (1), we may obtain by summation

$$(5)\quad E(\sum x^s f) = \sum x^s f + {}_sC_2 \cdot S_2 \cdot \sum x^{s-2} f + {}_sC_4 \cdot S_4 \sum x^{s-4} f + \cdots .$$

To illustrate: if we desire to shorten the distribution of table 1 by forming class intervals of dimension 9,

$$S_2 = \frac{1}{12}(9^2 - 1) = \frac{20}{3}, \qquad S_4 = \frac{1}{240}(9^2 - 1)(3 \cdot 9^2 - 7) = \frac{236}{3},$$

and by formula (5),

$$E(\sum x\ f) = \sum x f = -129$$

$$E(\sum x^2 f) = \sum x^2 f + {}_2C_2 \cdot S_2 \cdot \sum f = 77591 + \frac{20}{3} \cdot 244 = 79217^{2/3}$$

$$E(\sum x^3 f) = \sum x^3 f + {}_3C_2 \cdot S_2 \cdot \sum x f = -52005 + 3 \cdot \frac{20}{3}(-129) = -54585$$

$$E(\sum x^4 f) = \sum x^4 f + {}_4C_2 \cdot S_2 \cdot \sum x^2 f + {}_4C_4 \cdot S_4 \cdot \sum f$$

$$= 69239951 + 6 \cdot \frac{20}{3} \cdot 77591 + \frac{236}{3} \cdot 244 = 72362785^{2/3}.$$

Since these expected values are identical with those computed directly in table 3, we see that formula (5) provides the adjustments necessary to eliminate the effect of the systematic errors caused by grouping.

Dividing both sides of (5) by Σf yields

$$(6)\quad E(\mu'_s) = \mu'_s + {}_sC_2 \cdot S_2 \cdot \mu'_{s-2} + {}_sC_4 \cdot S_4 \mu'_{s-4} + {}_sC_6 \cdot S_6 \cdot \mu'_{s-6} + \cdots ,$$

that is

$$E(\mu'_1) = \mu'_1$$

$$E(\mu'_2) = \mu'_2 + \frac{1}{12}(k^2 - 1)$$

$$E(\mu_3') = \mu_3' + \frac{3}{12}(k^2 - 1)\,\mu_1'$$

$$E(\mu_4') = \mu_4' + \frac{6}{12}(k^2 - 1)\,\mu_2' + \frac{1}{240}(k^2 - 1)(3k^2 - 7)$$

$$E(\mu_5') = \mu_5' + \frac{10}{12}(k^2 - 1)\,\mu_3' + \frac{5}{240}(k^2 - 1)(3k^2 - 7)\,\mu_1'$$

etc.

In numerical computations we generally prefer to select the class interval as the unit of x and in this case we have

$$E(\mu_1') = \mu_1'$$

$$E(\mu_2') = \mu_2' + \frac{1}{12}\left(1 - \frac{1}{k^2}\right)$$

$$E(\mu_3') = \mu_3' + \frac{3}{12}\left(1 - \frac{1}{k^2}\right)\mu_1'$$

$$E(\mu_4') = \mu_4' + \frac{6}{12}\left(1 - \frac{1}{k^2}\right)\mu_2' + \frac{1}{240}\left(1 - \frac{1}{k^2}\right)\left(3 - \frac{7}{k^2}\right)$$

etc.

Ordinarily we are interested in estimating the values of the moments that would have been obtained if we had not used the time-saving device of grouping the variates and therefore we solve the previous set of equations for the moments of the ungrouped distribution and obtain

$$(7)\qquad \begin{cases} \mu_1' = E(\mu_1') \\ \mu_2' = E(\mu_2') - \dfrac{1}{12}\left(1 - \dfrac{1}{k^2}\right) \\ \mu_3' = E(\mu_3') - \dfrac{3}{12}\left(1 - \dfrac{1}{k^2}\right)E(\mu_1') \\ \mu_4' = E(\mu_4') - \dfrac{6}{12}\left(1 - \dfrac{1}{k^2}\right)E(\mu_2') + \dfrac{1}{240}\left(1 - \dfrac{1}{k^2}\right)\left(7 - \dfrac{3}{k^2}\right) \\ \qquad\qquad \text{etc.} \end{cases}$$

In general we may write, corresponding to formula (6),

$$(8)\qquad \mu_s' = E(\mu_s') - {}_sC_2 \cdot P_2 \cdot E(\mu_{s-2}') + {}_sC_4 \cdot P_4 \cdot E(\mu_{s-4}') - \cdots$$

where

$$P_2 = \frac{1}{12}\left(1 - \frac{1}{k^2}\right)$$

$$P_4 = \frac{1}{240}\left(1 - \frac{1}{k^2}\right)\left(7 - \frac{3}{k^2}\right)$$

$$P_6 = \frac{1}{1344}\left(1 - \frac{1}{k^2}\right)\left(31 - \frac{18}{k^2} + \frac{3}{k^4}\right)$$

$$P_8 = \frac{1}{11520}\left(1 - \frac{1}{k^2}\right)\left(381 - \frac{239}{k^2} + \frac{55}{k^4} - \frac{5}{k^6}\right)$$

$$P_{10} = \frac{1}{33792}\left(1 - \frac{1}{k^2}\right)\left(2555 - \frac{1636}{k^2} + \frac{410}{k^4} - \frac{52}{k^6} + \frac{3}{k^8}\right)$$

$$P_{12} = \frac{1}{5591040}\left(1 - \frac{1}{k^2}\right)\left(1414477 - \frac{910573}{k^2} + \frac{233570}{k^4} - \frac{32410}{k^6} + \frac{2625}{k^8} - \frac{105}{k^{10}}\right).$$

In actual problems we do not know the exact values of the expectations involved in formulae (7) and (8), and are forced to obtain mere approximations by utilizing in their stead the corresponding moments computed from the single chance grouped distribution. These approximations correspond to those employed in the theory of probable error, namely, substitutions of the moments derived from a single sample for the corresponding expected moments of the parent population.

The adjustments so far considered may properly be referred to as *Sheppard's adjustments about a fixed point.* At first thought it might appear that we might obtain corresponding formulae for the expectations of moments *about the mean* by merely dropping the primes in formula (6) and obtain, for example,

$$\mu_2 = E(\mu_2) - \frac{1}{12}(k^2 - 1),$$

but unfortunately this is not true. For example, the exact value for the variance of the distribution of table 1 is $18915563/244^2$. Using the summations of table 3 and computing the variance for each of the nine groupings yields

$$E(\mu_2) = \frac{1}{9.244^2}[18791595 + 19098180 + 18963315 + 19438947 + 19981080 + 19547811 + 19590984 + 19159011 + 19217736] \tag{9}$$

$$= 19309851/244^2.$$

Since $\frac{1}{12}(k^2 - 1) = \frac{1}{12}(9^2 - 1) = 20/3$ we see that

$$\mu_2 < E(\mu_2) - \frac{1}{12}(k^2 - 1).$$

In the theory of sampling we differentiate between the standard errors of moments about a fixed point and the standard error of moments about the mean

of the sample. Apparently writers on the subject of Sheppard's adjustments have overlooked the case of adjustments about the mean, although the solution for the second moment is readily obtained as follows:

$$E(\mu_2) = E(\mu_2' - M^2) = E(\mu_2') - E(M^2)$$
$$= \mu_2' + \frac{1}{12}(k^2 - 1) - \frac{1}{k}(M_1^2 + M_2^2 + \cdots + M_k^2),$$

where M_i represents the mean of the i-th of the k different grouped distributions. Since

$$\mu_2 = \mu_2' - M^2 = \mu_2' - \frac{1}{k}(M_1 + M_2 + \cdots + M_k),$$

$$E(\mu_2) = \mu_2 + \frac{1}{12}(k^2 - 1)$$
$$- \left[\frac{M_1^2 + M_2^2 + \cdots + M_k^2}{k} - \left(\frac{M_1 + M_2 + \cdots + M_k}{k}\right)^2\right].$$

But since for any set of k variates

$$\sigma_\nu^2 = \frac{\Sigma\nu^2}{k} - \left(\frac{\Sigma\nu}{k}\right)^2,$$

we have that

$$E(\mu_2) = \mu_2 + \frac{1}{12}(k^2 - 1) - \sigma_M^2. \tag{10}$$

Referring back to table 3 we find that

$$\sigma_M^2 = \frac{7856}{3.(244^2)}$$

and the numerical results now satisfy equation (10).

For the benefit of those interested in unsolved problems of mathematical statistics we may say that nothing appears to have been written as yet on the most important problem associated with the systematic errors due to grouping. It is of course desirable to eliminate these systematic errors introduced by grouping, but it is even more important to investigate the distribution of the accidental errors that remain after the systematic errors have been eliminated. For example it is gratifying to know that no systematic errors are present in the Σxf column of table 3 and that equation (6) will enable us to add a constant to each summation of the $\Sigma x^3 f$ column so that the mean of these adjusted values will agree with the value $\Sigma x^3 f = -52005$ obtained in table 1. It is rather disconcerting, however, to realize that in actual practice we *may* in the case of discrete variates and *must* in the case of continuous variates select an arbitrary set of class limits for our recorded data, and that after adjustments for grouping

have been made, our estimates of the true values of the moments of the distribution will—as in table 3—depend so much upon the choice of these limits. Thus, the standard error of the mean attributed to grouping is

$$\sigma_M = \frac{1}{244}\sqrt{\frac{7856}{3}} = 0.21,$$

which is about twenty percent as large as the approximation for the standard error of the mean due to sampling from an infinite parent population, namely,

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = 1.15.$$

If one will take the trouble to compute the values of μ_3 and μ_4 for each of the distributions of table 2, utilizing the summations of table 3, and then compute and compare the values of σ_{μ_3} and σ_{μ_4} due to grouping with the corresponding functions associated with sampling, he will realize the seriousness of the situation.

Summary

The formula for Sheppard's adjustments for distributions of grouped discrete variates was first given without proof in the Editorial of Vol. 1, No. 1 of the *Annals* (page 111). The method used to develop the general formula was extremely laborious and paralleled the method used for the case of continuous variates in the *Handbook of Mathematical Statistics*, Chapter 7, except that the calculus of finite differences was employed. A more satisfactory proof of this formula was presented by Dr. J. R. Abernethy in Vol. 4, No. 4 of the *Annals* in an article entitled *"On the Elimination of Systematic Errors Due to Grouping."* An extremely elegant development of the same formula and an extension to the case of two variables appears elsewhere in this volume by Professor C. C. Craig. From the point of view of expectations, all of these developments are adjustments about a fixed point, although this fixed point may be selected arbitrarily at the mean of the distribution in question. The obtaining of formulae for the adjustments about the mean of each grouping and the distribution of the accidental errors that remain after these systematic errors have been removed has apparently been neglected to date and should interest students of mathematical statistics.

From a mathematical standpoint, the development of this paper is the simplest of all that have appeared to date: the adjustments for the first four moments can be worked out with the aid of the binomial considerations leading to formula (3) and the following well known formulae for the sums of the powers of the first n integers:

$$S_1 = \frac{n(n+1)}{2} \qquad S_3 = \frac{n^2(n+1)^2}{4}$$

$$S_2 = \frac{n(n+1)(2n+1)}{6} \qquad S_4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}.$$

One should note that the condition of high contact is not required in this paper or in the developments of Abernethy or Craig. The results of the three preceding papers agree with those obtained about a fixed point in this paper, but fail to hold for the case of expectations about the mean, if we accept the following definition:

$$E(\mu_s) = \frac{1}{k}(\mu_{s:1} + \mu_{s:2} + \cdots + \mu_{s:k}), \qquad (s = 2, 3, \cdots)$$

where $\mu_{s:i}$ designates the s-th moment computed about the mean of the i-th grouped distribution, $(1 \leqq i \leqq k)$.

H. C. CARVER.

ON A GENERAL SOLUTION FOR THE PARAMETERS OF ANY FUNCTION WITH APPLICATION TO THE THEORY OF ORGANIC GROWTH

By Harry Sylvester Will

Part I

I. **The Problem Stated.** A type of problem which continually arises in the ordinary course of statistical analysis is that of determining the numerical values of the parameters of a function used to represent a series of observational data. In mathematical terminology, the problem may be stated as follows:

Given, the observational series $Y_0, Y_1, \cdots Y_{n-1}$.

Assumed, the function $y = f(x, a, b, c, \cdots)$.

To find, the numerical values of the parameters $a, b, c, \cdots$.

If the function $f(x, a, b, c, \cdots)$ is linear in the parameters, the desired solution is easily obtained by familiar methods. In cases where the function is not linear, the standard procedure is to reduce it to the linear form by expansion into Taylor's series, thus:

$$f(x, a, b, c) = f(x, a_0b_0c_0) + f_a(x, a_0b_0c_0)\cdot\Delta a + f_b(x, a_0b_0c_0)\cdot\Delta b + f_c(x, a_0b_0c_0)\cdot\Delta c, \tag{1}$$

where $a = a_0 + \Delta a$, $b = b_0 + \Delta b$, $c = c_0 + \Delta c$.

The use of this method suffers from the excessive labor involved as the number of parameters to be determined increases. In cases where satisfactory values of the first approximations $a_0b_0c_0$ are not obtainable, the solution becomes impossible. The basic difficulty arises from the consideration that the Taylor theorem requires that the increments Δa, Δb, Δc shall be very small quantities.

A method of successive approximation which makes feasible the reduction of gross errors in the corrections will, I take it, be of considerable interest to mathematical statisticians. Let us, therefore, proceed to the development of a technique which accomplishes precisely this result.

II. **The Theta Technique.** Let us begin our development with the following restatement of the technical problem involved:

Given, the observational series $Y_0, Y_1, \cdots Y_{n-1}$.

Assumed, the function $y = f(x, (a_0 + \theta_1\Delta a), (b_0 + \theta_2\Delta b), (c_0 + \theta_3\Delta c))$.

To find, the values of $\theta_1, \theta_2, \theta_3$.

In this set of relations, a_0, b_0, c_0 and $\Delta a, \Delta b, \Delta c$ are known quantities; while θ_1, θ_2 and θ_3 are each assumed not to exceed ± 1 in value. It follows, therefore,

that the adjusted values of a, b, and c lie within the bounds $a_0 \pm \Delta a$, $b_0 \pm \Delta b$, $c_0 \pm \Delta c$. We may, then, write the following:

$$\begin{aligned} a_1 &= a_0 - \Delta a; \qquad a_2 = a_0 + \Delta a. \\ b_1 &= b_0 - \Delta b; \qquad b_2 = b_0 + \Delta b. \\ c_1 &= c_0 - \Delta c; \qquad c_2 = c_0 + \Delta c. \end{aligned} \tag{2}$$

The values of θ_1, θ_2 and θ_3 are determined by the following procedure:

First, form the function y from all possible combinations of a_1a_2, b_1b_2, c_1c_2, thus:

$$\begin{aligned} y_{111} &= f(x, a_1b_1c_1). \\ y_{112} &= f(x, a_1b_1c_2). \\ &\cdots\cdots\cdots\cdots \\ y_{222} &= f(x, a_2b_2c_2). \end{aligned} \tag{3}$$

In the case of p parameters, we can evidently form 2^p distinct sets of n values for the function y_{iii}. Since the assigned values of parameters are mere approximations to their true values, each computed set of values for the function y_{iii} will differ from the true values $y = f(x, abc)$.

Second, form the theoretical residuals $y_{iii} - y$, and then compute the corresponding standard errors of estimate σ_{iii}. There will, accordingly, be 2^p values of σ determined, each value being a measure of the error committed in assuming the corresponding approximations to parameters; thus, σ_{111} measures the errors committed in assuming the combination $a_1b_1c_1$; σ_{112} measures the errors committed in assuming $a_1b_1c_2$; $\cdots$; σ_{222} measures the errors committed in assuming $a_2b_2c_2$.

Third, taking the squared reciprocal of σ as a measure of the reliability of a given determination of y_{iii} from the parameters $a_ib_ic_i$, we may form the following comparative tests of the reliability of the 2^p sets of the values of y_{iii}, thus:

$$\begin{aligned} \omega_{111} &= \sigma_{111}^{-2} : (\sigma_{111}^{-2} + \sigma_{112}^{-2} + \cdots + \sigma_{222}^{-2}) = \sigma_{111}^{-2} : \textstyle\sum \sigma_{iii}^{-2}. \\ \omega_{112} &= \sigma_{112}^{-2} : \textstyle\sum \sigma_{iii}^{-2}. \\ &\cdots\cdots\cdots\cdots\cdots \\ \omega_{222} &= \sigma_{222}^{-2} : \textstyle\sum \sigma_{iii}^{-2}. \end{aligned} \tag{4}$$

Omega, we shall term the *test constant*. Obviously, $\Sigma\omega_{iii} = 1$.

Fourth, assuming three parameters, let us tabulate the possible subscripts of omega according to the following scheme:

$\omega(a_1)$	$\omega(a_2)$	$\omega(b_1)$	$\omega(b_2)$	$\omega(c_1)$	$\omega(c_2)$
111	211	111	121	111	112
121	221	211	221	211	212
112	212	112	122	121	122
122	222	212	222	221	222

In this table, the subscripts are in the order of *abc*; so that 111 denotes $\omega(a_1b_1c_1)$; 112 denotes $\omega(a_1b_1c_2)$; etc. Comparing columns $\omega(a_1)$ and $\omega(a_2)$, we observe that the *bc* subscripts are identical for both; while the a_1 subscripts of the first column are replaced by the a_2 subscripts in the second column. Again, comparing columns $\omega(b_1)$ and $\omega(b_2)$, we see that the *ac* subscripts are identical for both; while the b_1 subscripts of the one column are replaced by the b_2 subscripts in the other. Finally, comparing columns $\omega(c_1)$ and $\omega(c_2)$, we note that the *ab* scripts are identical for both; while the c_1 subscripts of the one column are replaced by c_2 subscripts in the other.

Fifth, let us form the column summations $\Sigma\omega(a_1)$, $\Sigma\omega(a_2)$; $\Sigma\omega(b_1)$, $\Sigma\omega(b_2)$; and $\Sigma\omega(c_1)$, $\Sigma\omega(c_2)$. Since the columns $\omega(a_1)$ and $\omega(a_2)$ differ only with respect to the α subscripts, the difference in value between the sums $\Sigma\omega(a_1)$ and $\Sigma\omega(a_2)$ can be due to differences in value between a_1 and a_2 only, and are not at all affected by differences in value between b_1b_2 and c_1c_2. $\Sigma\omega(a_1)$ and $\Sigma\omega(a_2)$ may, therefore, be regarded as the weights of a_1 and a_2 to be used in determining the adjusted value of a; for $\Sigma\omega(a_1) + \Sigma\omega(a_2) = 1$.

We may, then, write the following relations:

$$\begin{aligned} a &= \Sigma\omega(a_1)\cdot a_1 + \Sigma\omega(a_2)\cdot a_2 = \Sigma\omega(a_1)\cdot(a_0 - \Delta a) + \Sigma\omega(a_2)\cdot(a_0 + \Delta a) \\ &= (\Sigma\omega(a_1) + \Sigma\omega(a_2))\cdot a_0 + (\Sigma\omega(a_2) - \Sigma\omega(a_1))\cdot\Delta a = a_0 + \theta(a)\cdot\Delta a. \end{aligned} \tag{5}$$

Since precisely similar reasoning applies to the parameters b_1, b_2 and c_1, c_2, we have the following definitive formulas for computing the values of theta:

$$\begin{aligned} \theta(a) &= \Sigma\omega(a_2) - \Sigma\omega(a_1). \\ \theta(b) &= \Sigma\omega(b_2) - \Sigma\omega(b_1). \\ \theta(c) &= \Sigma\omega(c_2) - \Sigma\omega(c_1). \end{aligned} \tag{6}$$

As the adjusted values of parameters, we have:

$$\begin{aligned} a &= a_0 + \theta(a)\cdot\Delta a. \\ b &= b_0 + \theta(b)\cdot\Delta b. \\ c &= c_0 + \theta(c)\cdot\Delta c. \end{aligned} \tag{7}$$

In this development of the theta technique, we have determined $\sigma_{...}$ from the theoretical residuals $y_{...} - y$. This has served well the purposes of exposition; but, since the true values of the function y are unknown, we must, in practice, compute $\sigma_{...}$ from the observational residuals $y_{...} - Y$. Later in the memoir, it will be shown how the computation of θ may, in numerous cases, be considerably abridged.

Part II

III. **The Principle of Malthus.** Since a determination of the numerical parameters of a given function by means of the theta technique must, at best,

involve a considerable amount of computation, I have chosen for purposes of demonstration a problem which is of much interest in itself. This problem, we shall state in the form of two questions:

First, what is the most appropriate mathematical form of the law of organic growth?

Second, how may the parameters of the indicated function be computed?

Thomas R. Malthus, in his famous essay on *The Principle of Population Growth* assumed that the proportional growth of human populations is properly defined by the differential equation,

$$\frac{1}{p} \cdot \frac{dp}{dt} = b, \tag{8}$$

where p is the population under consideration, t is the measure of time, and b is the stable or geometric rate of growth.

This formula has been destructively criticised on the ground that it fails wholly to give a mathematical description of the manner in which population growth is kept within bounds. So far as any implication of the formula is concerned, populations may grow to infinite magnitudes. An attempt to represent growth by its use must, therefore, result in a succession of discontinuities which are incompatible with the observed facts of organic growth.

IV. **The Symmetric Logistic.** In three memoirs published in 1838, 1845 and 1847, it was suggested by M. Verhulst, Professor of Mathematics in the Ecole Militaire in Brussels, that the rate of population growth might be stated as a function of the population itself. Assuming the limiting value of p to be H, this conception of the growth rate Verhulst expressed by the differential equation,

$$\frac{1}{p} \cdot \frac{dp}{dt} = -b(1 - pH^{-1}). \tag{9}$$

Since this equation expresses proportional growth as a linear function of p, it is the simplest relation of its kind that may be conceived. In representing the rate of growth as a quantity which approaches zero as the population approaches its limiting value, it makes, indeed, a significant advance over the Malthusian formula. Nevertheless, the equation is subject to an interesting limitation, the nature of which is made evident by an examination of the integral form of the function, namely:

$$p = H:[1 + e^{a+bt}]. \tag{10}$$

This we shall now prove to be rotationally symmetric with respect to the point of inflection.

Differentiating equation (9) a second time, we have,

$$\begin{aligned} d^2p &= -b\, dp[p(1 - H^{-1}p)]dt \\ &= p[p^{-2}dp^2 - b\, d^2t + bH^{-1}p\, d^2t + bH^{-1}dp\, dt] \\ &= p^{-1}dp^2 + bH^{-1}p\, dp\, dt. \end{aligned}$$

Hence,

$$\frac{d^2p}{dt^2} = b^2p(1 - H^{-1}p)^2 - b^2H^{-1}p^2(1 - H^{-1}p).$$

Setting $\frac{d^2p}{dt^2} = 0$, we get,

$$1 - 2H^{-1}p = 0.$$

Or

$$p = H/2, \tag{11}$$

which gives the value of p at the point of inflection.

Substituting for p from (10), and solving for t, we have,

$$t_i = -a/b, \tag{12}$$

where t_i is the point of inflection of the function p.

Denoting the magnitude of the population at time t_i by p_i, its magnitude at time t_{i+k} by p_{i+k}, and its magnitude at the time t_{i-k} by p_{i-k}, we have,

$$p_i = H:[1 + e^{a+b(-a/b)}] = H/2. \tag{13}$$

$$p_{i+k} = H:[1 + e^{a+b(t+k\Delta t)}] = H:[1 + e^{bk\Delta t}]. \tag{14}$$

$$p_{i-k} = H:[1 + e^{a+b(t-k\Delta t)}] = H:[1 + e^{-bk\Delta t}]. \tag{15}$$

Measuring p in units of H and setting $u = e^{bk\Delta t}$, we may rewrite these last three equations as follows:

$$H^{-1}p_i = 1/2.$$

$$H^{-1}p_{i+k} = 1:[1 + u].$$

$$H^{-1}p_{i-k} = 1:[1 + u^{-1}].$$

On the hypothesis of rotational symmetry, we have, by subtraction,

$$H^{-1}p_{i+k} - 1/2 = 1/2 - H^{-1}p_{i-k}.$$

In proof, we have:

$$\begin{aligned} 1:[1 + u] &= 1 - 1:[1 + u^{-1}] \\ &= u^{-1}:[1 + u^{-1}] \\ &= 1:[u + 1]. \end{aligned}$$

q. e. d.

Part III

V. **Criticisms of the Logistic.** Because of its symmetric form, many critics have called into question the finality of the logistic as a universal repre-

sentation of population growth. That it applies in particular cases, they contend, is no reason for holding that it must apply in general. Professors Raymond Pearl and Lowell J. Reed of Johns Hopkins University—to whom we are indebted for the rediscovery of the earlier researches of Verhulst—have proposed, as the proper form of the generalized growth curve, the following function:

$$p = H:[1 + e^{a+bt+ct^2+dt^3}]. \tag{16}$$

In their view, this equation is suited not only to representing a single cycle of growth, but two successive cycles as well. This claim, however, must be rejected; for, if true, it would mean that one cycle of growth is predictable from another, a circumstance which is clearly inconsistent with the assumptions laid down by these same investigators.

Moreover, so far as I can learn from their published writings, these authors have never considered the implications of the differential form of the function they propose.

Differentiating (16), we have,

$$\frac{1}{p} \cdot \frac{dp}{dt} = -(b + 2cx + 3dx^2)(1 - H^{-1}p).$$

Here, we find the stable growth constant of Malthus replaced by an expression which is quadratic in t. This means that, for a population which is freed of a restraining limit, proportional growth tends generally toward infinite values. If there are any facts to support such a conception of organic growth, I do not know what they are, and must, perforce, reject the contention that equation (16) is the generalized form of the Verhulst function.

VI. **Fundamental Assumptions.** In order to represent the phenomenon of population growth mathematically, I hold the following assumptions to be necessary:

(a) Under favoring conditions, population may increase at a constant geometric rate.

(b) Under all circumstances, the rate of growth must be a finite and continuous quantity.

(c) The magnitude of a population is always a positive, real number.

(d) The growth of population tends toward restriction within definite bounds.

(e) The growth of population is a function of time.

(f) The basic conditions of growth are free of cataclysmic disturbances.

The first of these assumptions is given in recognition of well known facts concerning organic growth. The second is necessary because, even when the size of a population is freed of definite restriction, the pattern of growth is not necessarily geometric. The third assumption affirms the absurdity of representing a population as a negative or infinite quantity. The fourth merely asserts the indisputable fact that the organism must always grow in a finite environment. The fifth gives place to the concept of growth as the resultant of a complex of

causes, no one of which can be isolated as an entirely independent variable. While the final assumption recognizes that major disturbing influences may profoundly affect the course of growth.

VII. **The Skew Logistic.** In accord with our fundamental assumptions, we may form the following differential equations:

$$\frac{1}{p'} \cdot \frac{dp'}{dt} = -[b + sm \cdot \cos(m(t+q))][1 - H^{-1}p'] \qquad \text{Type } \alpha$$

$$= -[b + 2sm^2(t+q) : (1 + m^2(t+q)^2)][1 - H^{-1}p'] \qquad \text{Type } \beta \quad (17)$$

$$= -[b + sm^2(t+q) : \sqrt{1 + m^2(t+q)^2}][1 - H^{-1}p'] \qquad \text{Type } \gamma$$

In these equations, $p' = p - L$, and measures p from its lower limit as origin. On separating variables, the following integrations may be performed:

$$-\int [dp' : (p'(1 - H^{-1}p'))] = -\log[p' : (1 - H^{-1}p')] = \log[(H - p') : (Hp')].$$

Writing $z = m(t+q)$, $dz = mdt$; so that we have:

$$b\int dt + s\int \cos z\, dz = A + bt + s \cdot \sin z.$$

$$b\int dt + 2s\int [z : (1 + z^2)]dz = A + bt + s \cdot \log(1 + z^2).$$

$$b\int dt + s\int [z : \sqrt{1 + z^2}]dz = A + bt + s\sqrt{1 + z^2}.$$

From these integrals, we form the following equations:

$$\log[(H - p') : (Hp')] = A + bt + s \cdot \sin[m(t+q)].$$

$$\log[(H - p') : (Hp')] = A + bt + s \cdot \log[1 + m^2(t+q)^2].$$

$$\log[(H - p') : (Hp')] = A + bt + s \cdot \sqrt{1 + m^2(t+q)^2}.$$

We have, finally, on taking antilogarithms and making the substitutions $p = p' + L$, $a = A - \log H$:

$$p = L + H : [1 + e^{a+bt+s \cdot \sin(m(t+q))}]. \qquad \text{Type } \alpha$$

$$p = L + H : [1 + e^{a+bt+s \cdot \log(1+m^2(t+q)^2)}]. \qquad \text{Type } \beta \quad (18)$$

$$p = L + H : [1 + e^{a+bt+s\sqrt{1+m^2(t+q)^2}}]. \qquad \text{Type } \gamma$$

These equations give the normal forms of the skew logistic.

VIII. **Properties of the Skew Logistic.** We may deduce the properties of the skew logistic by examining both its differential and integral forms. Considering the derivative of Type α, we note that the Malthusian constant b is

replaced by a trigonometric function whose amplitude is $b \pm sm$, and whose phase depends on the values of m and q. When $b \pm sm = 0$, the derivative must also equal zero, and a flat point in the curve of p is indicated. When b is absolutely less than sm, the derivative changes sign and the curve of p reverses its direction. Thus, the integral form of Type α modifies the symmetric form of the logistic by a succession of minor cycles in which the rate of growth is alternately accelerated and retarded.

Considering Type β, we find the Malthusian constant replaced by a function whose maximum and minimum values are attained when $t = m^{-1} - q$. Obviously, therefore, this function passes through a single period whose amplitude is $b \pm sm$, and whose phases are b, $b + sm$, b, $b - sm$, b. When $b \pm sm = 0$, a flat point in the curve of p is generated. The effect of skewness on the rate of growth passes through two double phases. Where b and s are of the same sign, these phases are: first, increasing retardation followed by decreasing retardation when $t + q$ is negative; and, second, increasing acceleration followed by decreasing acceleration when $t + q$ is positive. Where b and s are of opposite sign, the corresponding phases are: first, increasing acceleration followed by decreasing acceleration when $t + q$ is negative; and, second, increasing retardation followed by decreasing retardation when $t + q$ is positive. It is to be noted that, when sm is absolutely greater than b, the derivative will change sign twice before the upper limit is reached. Under these circumstances, the function p passes through a double reversal of direction.

Considering Type γ, we find the Malthusian constant of the derivative replaced by a function which is aperiodic and which approaches the limits $b \pm sm$ as t approaches $\pm \infty$. When b and s are of the same sign, skewness passes through the two following phases: first, the phase of decreasing retardation when $t + q$ is negative; and, second, the phase of increasing acceleration when $t + q$ is positive. On the other hand, when b and s are of opposite sign, the corresponding phases are: first, that of decreasing acceleration when $t + q$ is negative; and, second, that of increasing retardation when $t + q$ is positive. When sm is absolutely greater than b, the derivative changes sign, and the function p passes from a continuously increasing phase to a continuously decreasing phase, or *vice versa*.

In general, it may be said of all three types—α, β and γ—that, if the derivative is not restricted to a single change of sign, L denotes a lower asymptote of the function p; while, under the same conditions, H denotes the higher limit approached by the function $p - L$. When H is negative, the effect is to make L an upper, and $L - H$ a lower, asymptote of the curve p.

In the case of Type γ, when the function p makes a single change of sign, either H or L becomes a maximum (or minimum) value instead of an asymptote of the curve. In this event, it will be noted that the factor $1 - H^{-1}p$ appearing in the derivative does not approach zero as a limit with increasing values of t, but rather passes through a minimum and then approaches the limit 1 in either direction.

The parameter s may be positive or negative in sign, and is termed the index of skewness or, briefly, the *skewness* of the function. Obviously, m is always positive, and, since it determines the rate at which skewness develops, is properly termed the *development*. The point in time at which skewness passes from an accelerating to a retarding phase, or *vice versa*, is fixed by the value of q, which is, therefore, termed the *transition*. The parameter b, as has already been stated, is termed the stable growth tendency or, technically, the *stability* of the function. And since the position of the curve p on an arbitrary time scale will vary with the value of a, this parameter I have designated the *location*.

In all three types of the skew logistic, if $e^{\psi(t)}$ is a continuously decreasing function and both H and L are positive, the curve of p may be described as of the *rising hillside form*. In the case of Type γ, if the derivative changes from positive to negative sign, the curve may be described as *mountain formed*. If $e^{\psi(t)}$ increases continuously, the curve is of the *falling hillside* variety, except when the derivative of Type γ changes from negative to positive sign, in which event a *valley form* is generated.

Part IV

IX. **Parameters of the Symmetric Logistic.** The numerical parameters of the symmetric logistic (10) are most easily determined by the method of differences. First, we write,

$$p_i^{-1} = C + e^{A+bt}, \tag{19}$$

where $C = H^{-1}$; $A = a - \log H$; and $i = 0, 1, 2, \cdots n - 1$.

Assuming Δt constant, let us give to t the increment $k\Delta t$, thus:

$$p_{i+k}^{-1} = C + e^{A+b(t+k\Delta t)}. \tag{20}$$

Subtracting (19) from (20), we obtain

$$\Delta_k p_i^{-1} = e^{A+b(t+k\Delta\cdot)} - e^{A+bt} = Be^{A+bt}, \tag{21}$$

where $B = e^{bk\Delta t} - 1$. The quantity $\Delta_k p_i^{-1} = p_{i+k}^{-1} - p_i^{-1}$ is termed a first order difference of rank k.

Giving to t in equation (21) the increment $k\Delta t$, we get

$$\Delta_k p_{i+k}^{-1} = Be^{A+b(t+k\Delta t)}. \tag{22}$$

Dividing (22) by (21), we have,

$$\Delta_k p_{i+k}^{-1} : \Delta_k p_i^{-1} = e^{bk\Delta t}.$$

Taking logarithms, we obtain

$$\Delta_k \log \Delta_k p_i^{-1} = \log \Delta_k p_{i+k}^{-1} - \log \Delta_k p_i^{-1} = bk\Delta t,$$

which defines the parameter b. We can form $n - 2k$ such equations. Hence,

b is uniquely determined by the relation

$$b = [\sum_{i=0}^{i=n-2k-1} \Delta_k \log \Delta_k P_i^{-1}] : [k(n - 2k)\Delta t]$$
$$= [\sum_{i=k}^{i=n-k-1} \log \Delta_k P_i^{-1} - \sum_{i=0}^{i=n-2k-1} \log \Delta_k P_i^{-1}] : [k(n + 2k)\Delta t], \quad (23)$$

where $k = n{:}3$ to the nearest integer.

Returning to (21), we have the following relation determining the value of A:

$$A = \log [\sum_{i=0}^{i=n-k-1} \Delta_k P_i^{-1}] - \log [B \sum_{i=0}^{i=n-k-1} e^{bt}]$$
$$= \log [\sum_{i=k}^{i=n-1} P_i^{-1} - \sum_{i=0}^{i=n-k-1} P_i^{-1}] - \log [B \sum_{i=0}^{i=n-k-1} e^{bt}], \quad (24)$$

where $k = n{:}2$ to the nearest integer.

From equation (19), we have

$$C = [\sum_{i=0}^{i=n-1} P_i^{-1} - e^A \sum_{i=0}^{i=n-1} e^{bt}] : n. \quad (25)$$

The values of H and a are, obviously, given by

$$H = C^{-1}. \quad (26)$$

$$a = A + \log H. \quad (27)$$

In the relations defining b, A and C, the values of P must be obtained from the observations. In computing the values of k, the formula is:

$$k = n(r + 1)^{-1},$$

where n is the number of observations, and r denotes the order of reduction involved in the defining relation.

In my first treatment of the subject, I assumed that the value of k for all orders of reduction might be determined from the reduction of highest order involved; but I have since found that I erred in this view. The point is that the function $\psi(p) = k^r(n - rk)$, discussed in the original memoir, must be maximized with respect to k separately for each order of difference involved; or, in other words, the rank constant k must be given a separate determination for each parameter defined if the most accurate results are to be obtained.

X. **Parameters of the Skew Logistic.** I shall now show how the method of differences may be used to abridge the computations involved in applying the theta technique to the determination of the parameters of the skew logistic. In this, as in the preceding section, we assume Δt constant.

Operating on Type γ of equation (18), we write

$$p_i = L + H : [1 + e^{a+bt+s\sqrt{1+m^2(t+q)^2}}]. \quad (28)$$

To begin with, let us write the transformation of ordinate

$$G = \log [H(p - L)^{-1} - 1].$$

Also, let us write

$$F = \sqrt{1 + m^2(t + q)^2}.$$

We may now rewrite equation (28) in the form

$$G_i = a + bt + sF_i. \tag{29}$$

Giving to t the increment $k\Delta t$, we have

$$G_{i+k} = a + b(t + k\Delta t) + sF_{i+k}. \tag{30}$$

Subtracting (29) from (30), we have,

$$\Delta_k G_i = bk\Delta t + s\Delta_k F_i. \tag{31}$$

Again giving to t the increment $k\Delta t$, we obtain

$$\Delta_k G_{i+k} = bk\Delta(t + k\Delta t) + s\Delta_k F_{i+k}. \tag{32}$$

Subtracting (31) from (32), we obtain

$$\Delta_k G_{i+k} - \Delta_k G_i = (bk\Delta t - bk\Delta t) + s(\Delta_k F_{i+k} - \Delta_k F_i),$$

or

$$\Delta_k^2 G_i = s\Delta_k^2 F_i. \tag{33}$$

We can form $n - 2k$ such equations, and may, therefore, form $n - 2k$ approximations to the value of the parameter s, as follows:

$$s_i = [\Delta_k^2 G_i]:[\Delta_k^2 F_i]; \qquad i = 0, 1, \cdots, n - 2k - 1.$$

Taking the mean value of the set s_i as its most probable value, we have,

$$s_0(HL \cdot mq) = \Sigma s_i:(n - 2k);\ k = n:3 \text{ to the nearest integer} \tag{34}$$

In this determination of s_0, the only parameters directly involved are H, L, m and q, the parameters a and b having been eliminated. By assigning values to H_0, L_0, m_0 and q_0, we may, on setting up the arbitrary corrections ΔH, ΔL, Δm and Δq, write down the following:

$$H_1 = H_0 - \Delta H; \quad H_2 = H_0 + \Delta H; \quad L_1 = L_0 - \Delta L; \quad L_2 = L_0 + \Delta L;$$
$$m_1 = m_0 - \Delta m; \quad m_2 = m_0 + \Delta m; \quad q_1 = q_0 - \Delta q; \quad q_2 = q_0 + \Delta q.$$

Since s_0 is a function of H, L, m and q, we may, by entering the subscripts of the combination $HL \cdot mq$, tabulate the possible determinations of s_0 as follows:

11·11	11·12	11·21	11·22
12·11	12·12	12·21	12·22
21·11	21·12	21·21	21·22
22·11	22·12	22·21	22·22

In this tabulation, the subscripts of parameters are in the order of $HL \cdot mq$; so that 12·21 denotes $s_0(H_1L_2 \cdot m_2q_1)$, etc.

From the table, it is seen that we may compute $2^4 = 16$ distinct sets of approximations to $s_0(HL \cdot mq)$. Since the true values of H, L, m and q are unknown, each set of approximations s_i will show a characteristic variation about its mean

value, s_0. This variation is most conveniently measured by the mean deviation

$$\epsilon = (s_0 - s_0')2N':N = (s_0'' - s_0)2N'':N, \tag{35}$$

where the second relation serves as a check on the computation by the first; $N = n - 2k$; N' denotes the number of items s_i which are *less* than s_0 in value, and N'', the number of items s_i which are *greater* than s_0 in value; while s_0' denotes the mean of the N' values of s_i which are less than s_0, and s_0'', the mean of the N'' values of s_i which are greater than s_0.

The reliability of a given value of s_0 as a measure of the central tendency of the corresponding set s_i is sufficiently determined by ϵ^{-2}, which serves at the same time to measure the reliability of the combination $HLmq$ figuring in the computation of the given set s_i. We may, therefore, compute the values of the test constant, ω, directly from the values of ϵ^{-2} by means of the relation,

$$\omega(HL\cdot mq) = \epsilon_j^{-2}:[\epsilon_{11\cdot 11}^{-2} + \epsilon_{11\cdot 12}^{-2} + \cdots + \epsilon_{22\ 22}^{-2}] = \epsilon_j^{-2}:\Sigma\epsilon^{-2}, \tag{36}$$

where $j = 11\cdot 11, 11\cdot 12, \cdots, 22\cdot 22$; $\Sigma\omega = 1$.

Since four values of theta are to be determined, we must arrange the sixteen values of omega in four ways, as shown by the following tabulation of subscripts:

$\omega(H_1)$	$\omega(H_2)$	$\omega(L_1)$	$\omega(L_2)$	$\omega(m_1)$	$\omega(m_2)$	$\omega(q_1)$	$\omega(q_2)$
11·11	21·11	11·11	12·11	11·11	11·21	11·11	11·12
11·12	21·12	11·12	12·12	11·12	11·22	11·21	11·22
11·21	21·21	11·21	12·21	12·11	12·21	12·11	12·12
11·22	21·22	11·22	12·22	12·12	12·22	12·21	12·22
12·11	22·11	21·11	22·11	21·11	21·21	21·11	21·12
12·12	22·12	21·12	22·12	21·12	21·22	21·21	21·22
12·21	22·21	21·21	22·21	22·11	22·21	22·11	22·12
12·22	22·22	21·22	22·22	22·12	22·22	22·21	22·22

Knowing the values of omega, we have at once,

$$\theta(H) = \Sigma\omega(H_2) - \Sigma\omega(H_1); \qquad \theta(L) = \Sigma\omega(L_2) - \Sigma\omega(L_1);$$
$$\theta(m) = \Sigma\omega(m_2) - \Sigma\omega(m_1); \qquad \theta(q) = \Sigma\omega(q_2) - \Sigma\omega(q_1). \tag{37}$$
$$H = H_0 + \theta(H)\cdot\Delta H; \qquad L = L_0 + \theta(L)\cdot\Delta L;$$
$$m = m_0 + \theta(m)\cdot\Delta m; \qquad q = q_0 + \theta(q)\cdot\Delta q. \tag{38}$$

The process of adjustment should be repeated until errors in the parameters diminish to negligible proportions.

With H, L, m and q known to a sufficient approximation, we may form anew the functions $G(H, L, m, q)$ and $F(H, L, m, q)$. We can then write $n - 2k$ equations of form (33), viz.:

$$\Delta_k^2 G_i = s\Delta_k^2 F_i.$$

Summing these equations, we have,

$$\Sigma\Delta_k^2 G_i = s\Sigma\Delta_k^2 F_i, \tag{39}$$

where $\Sigma\Delta_k^2 G_i = \sum_{i=2k}^{i=n-1} G_i - 2\sum_{i=k}^{i=n-k-1} G_i + \sum_{i=0}^{i=n-2k-1} G_i$;

$$\Sigma\Delta_k^2 F_i = \sum_{i=2k}^{i=n-1} F_i - 2\sum_{i=k}^{i=n-k-1} F_i + \sum_{i=0}^{i=n-2k-1} F_i;$$

where $k = n{:}3$ to the nearest integer.

The approximate value of s is now obtained from the relation

$$s = [\sum_{i=0}^{i=n-2k-1} \Delta_k^2 G_i] : [\sum_{i=0}^{i=n-2k-1} \Delta_k^2 F_i]. \tag{40}$$

Returning to equation (31), we solve for $bk\Delta t$, obtaining,

$$bk\Delta t = \Delta_k G_i - s\Delta_k F_i.$$

Since we can form $n - k$ such equations, the approximate value of b is given by the relation

$$b = [\Sigma\Delta_k G_i - s\Sigma\Delta_k F_i] : [k(n-k)\Delta t] \tag{41}$$

$$= [(\sum_{i=k}^{i=n-1} G_i - \sum_{i=0}^{i=n-k-1} G_i) - s(\sum_{i=k}^{i=n-1} F_i - \sum_{i=0}^{i=n-k-1} F_i)] : [k(n-k)\Delta t],$$

where $k = n{:}2$ to the nearest integer.

From equation (29), we obtain the approximate value of a as follows:

$$a = [\sum_{i=0}^{i=n-1} G_i - b\sum_{i=0}^{i=n-1} t - s\sum_{i=0}^{i=n-1} F_i] : n. \tag{42}$$

Comparing the abridged method of computing the values of theta here outlined with the general procedure of section II, it will be seen that we have been able to reduce the number of values of omega which it is necessary to determine from $2^7 = 128$ to $2^4 = 16$. In cases where L may be assumed to equal zero, the number of values of omega which must be computed is further reduced to $2^3 = 8$.

Part V

XI. **Symmetric Parameters for the Population of the United States.** I have determined the numerical values of the parameters of both the symmetric and the skew forms of the logistic from the population figures for the United States given by the Bureau of the Census. The only departure in the data from the census figures consists in the interpolation of all items to June 1st as the date of observation. The values of the symmetric parameters are computed from the data of Table I, as follows.

Setting $k = 15 \div 3 = 5$, we have, by equation (23),

$$\sum_0^4 \Delta_5 \log \Delta_5 P_i^{-1} = \sum_5^9 \log \Delta_5 P_i^{-1} - \sum_0^4 \log \Delta_5 P_i^{-1}$$

$$= \bar{9}.71878n - \bar{5}.14555n = -3.42677.$$

TABLE I

Data for the Symmetric Logistic

i	P^{-1}	$\Delta_5 P^{-1}$	$\log \Delta_5 P^{-1}$	10^{bi}
0	0.25582	−0.19724	$\bar{1}.29500_n$	1.00000
1	0.18939	−0.14627	$\bar{1}.16516_n$	0.72934
2	0.13885	−0.10704	$\bar{1}.02955_n$	0.53193
3	0.10431	−0.07838	$\bar{2}.89421_n$	0.38796
4	0.07770	−0.05776	$\bar{2}.76163_n$	0.28295
5	0.05858	−0.04269	$\bar{2}.63033_n$	0.20637
6	0.04312	−0.02996	$\bar{2}.47654_n$	0.15051
7	0.03181	−0.02098	$\bar{2}.32181_n$	0.10978
8	0.02593	−0.01650	$\bar{2}.21748_n$	0.08006
9	0.01994	−0.01182	$\bar{2}.07262_n$	0.05839
10	0.01589			0.04259
11	0.01316			0.03106
12	0.01083			0.02265
13	0.00943			0.01652
14	0.00812			0.01205
Σ	1.00288	−0.70864	$\bar{1}\bar{4}.86433_n$	3.66216

TABLE II(A)

Data for the Skew Logistic

i	G_1	G_2	F_{11}	F_{12}	F_{21}	F_{22}
0	+ 1.67998	+ 1.71132	6.47765	3.35261	9.65194	4.90306
1	+ 1.54968	+ 1.57779	5.68859	2.60000	8.45931	3.73631
2	+ 1.40690	+ 1.43878	4.90306	1.88680	7.26911	2.60000
3	+ 1.27698	+ 1.30927	4.12311	1.28062	6.08276	1.56205
4	+ 1.14130	+ 1.17416	3.35261	1.00000	4.90306	1.00000
5	+ 1.00816	+ 1.04179	2.60000	1.28062	3.73631	1.56205
6	+ 0.85948	+ 0.89428	1.88680	1.88680	2.60000	2.60000
7	+ 0.70540	+ 0.74193	1.28062	2.60000	1.56205	3.73631
8	+ 0.59699	+ 0.63515	1.00000	3.35261	1.00000	4.90306
9	+ 0.44841	+ 0.48956	1.28062	4.12311	1.56205	6.08276
10	+ 0.30840	+ 0.35346	1.88680	4.90306	2.60000	7.26911
11	+ 0.17992	+ 0.22981	2.60000	5.68859	3.73631	8.45931
12	+ 0.02885	+ 0.08647	3.35261	6.47765	4.90306	9.65194
13	− 0.09590	− 0.02968	4.12311	7.26911	6.08276	10.84620
14	− 0.25808	− 0.17670	4.90306	8.06226	7.26911	12.04159
Σ	+10.83647	+11.47739	49.45864	55.76384	71.41783	80.95375

TABLE II(B)

Data for the Skew Logistic

i	$\Delta_5^2 G_1$	$\Delta_5^2 G_2$	$\Delta_5^2 F_{11}$	$\Delta_5^2 F_{12}$	$\Delta_5^2 F_{21}$	$\Delta_5^2 F_{22}$
0	−0.02794	−0.01880	3.16445	5.69443	4.77932	9.04807
1	+0.01064	+0.01904	4.51499	4.51499	6.99562	6.99562
2	+0.02495	+0.04139	5.69443	3.16445	9.04807	4.77932
3	−0.01290	+0.00929	6.24622	1.84451	10.16552	2.60213
4	−0.01360	+0.01834	5.69443	0.81604	9.04807	0.87607
Σ	−0.01885	+0.06926	25.31452	16.03442	40.03660	24.30121

We note that $k(n - 2k)\Delta t = 5(15\text{–}10)1 = 25$; hence,

$$b = -3.42677 \div 25 = -0.1370708.$$

Next, set $k = 15 \div 2 = 7$, to the nearest integer; then, by equation (24), we get

$$\sum_0^7 \Delta_7 P_i^{-1} = \sum_7^{14} P_i^{-1} - \sum_0^7 P_i^{-1} = 0.10330 - 0.86777 = -0.76447;$$

$$B = 10^{bk\Delta t} - 1 = 10^{-0.1370708\times 7} - 1 = -0.89022; \qquad \sum_0^7 10^{bt} = 3.39884.$$

Hence,

$$A = \log[-0.76447] - \log[-0.89022 \times 3.39884] = \bar{1}.4025324.$$

We have next

$$\sum_0^{14} P_i^{-1} = 1.00288; \qquad \sum_0^{14} 10^{bt} = 3.66216; \qquad 10^A = 0.25266.$$

By equation (25), then, we obtain

$$C = [1.00288 - 0.25266 \times 3.66216] \div 15 = 0.0051747.$$

By equation (26), we get

$$H = C^{-1} = 193.25.$$

Finally, by equation (27), we obtain

$$a = A + \log H = \bar{1}.4025324 + 2.2861136 = 1.68865.$$

The point of inflection of the curve is given by

$$t_i = -a{:}b = 1.68865 \div 0.1370708 = 12.319.$$

XII. **Skew Parameters for the Population of the United States.** Assuming $L = 0$, we form

$$H_1 = 198.0 - 7.0 = 191.0; \qquad H_2 = 198.0 + 7.0 = 205.0.$$

$$m_1 = 1.0 - 0.2 = 0.8; \qquad m_2 = 1.0 + 0.2 = 1.2.$$

$$q_1 = -6.0 - 2.0 = -8.0; \qquad q_2 = -6.0 + 2.0 = -4.0.$$

Next, the primary data of Tables II(a) and II(b) are computed. Setting $k = 15 \div 5 = 3$, $n - k$ values of the 2^3 sets of s, are determined and entered in Table III(a). The values of s_0, ϵ and ω for each set are computed by equations (34), (35) and (36).

In Table III(b), the several values of ω are arranged according to their association: first, with H_1, H_2; second, with m_1, m_2; and, third, with q_1, q_2. The column sums yield the weights $\Sigma\omega$. The values of θ and the adjusted values of parameters are computed by equations (37) and (38):

TABLE III(A)

Data for the Computation of Θ

t	$s(1\,11)$	$s(1\,12)$	$s(1\,21)$	$s(1\,22)$	$s(2\,11)$	$s(2\,12)$	$s(2\,21)$	$s(2\,22)$
0	−0 00883	−0 00491	−0 00585	−0 00309	−0 00594	−0 00330	−0 00393	−0 00208
1	+0 00236	+0 00236	+0 00152	+0 00152	+0 00422	+0 00422	+0 00272	+0 00272
2	+0 00438	+0 00788	+0 00276	+0 00522	+0 00727	+0 01308	+0 00457	+0 00866
3	−0 00207	−0 00699	−0 00127	−0 00496	+0 00149	+0 00504	+0 00091	+0 00357
4	−0 00239	−0 01667	−0 00150	−0 01552	+0 00322	+0 02247	+0 00203	+0 02093
Σ	−0 00655	−0 01832	−0 00434	−0 01683	+0 01026	+0 04151	+0 00630	+0 03380
s_0	−0 00131	−0 00366	−0 00087	−0 00337	+0 00205	+0 00830	+0 00126	+0 00676
ϵ	+0 00374	+0 00703	+0 00241	+0 00550	+0 00342	+0 00404	+0 00222	+0 00643
ω	+0 10624	+0 03012	+0 25711	+0 04919	+0 12708	+0 09137	+0 00290	+0 03061

TABLE III(B)

Data for the Computation of Θ

	$\omega(h_1)$	$\omega(h_2)$	$\omega(m_1)$	$\omega(m_2)$	$\omega(q_1)$	$\omega(q_2)$
	0 1062	0 1271	0 1062	0 2571	0 1062	0 0301
	0 0301	0 0914	0 0301	0 0492	0 2571	0 0492
	0 2571	0 3029	0 1271	0 3029	0 1271	0 0914
	0 0492	0 0360	0 0914	0 0360	0 3029	0 0360
Σ	0 4426	0 5574	0 3548	0 6452	0 7933	0 2067

TABLE IV(A)

Summary of Adjustments

Parameter	Estimated Value	Δ	Θ	$\Delta\cdot\Theta$	Adjusted Value
H	+198 0	+7 0	+0 1148	+0.8036	+198 80
m	+ 1.0	+0 2	+0 2904	+0 05808	+1 05808
q	− 6 0	+2 0	−0 5866	−1.1732	−7 1732

TABLE IV(B)

Final Transformations

i	$G(Hmq)$	$F(Hmq)$
0	1.69772	7.65559
1	1.56410	6.60800
2	1.42495	5.46440
3	1.29526	4.52752
4	1.15991	3.50336
5	1.02722	2.50753
6	0.87921	1.59408
7	0.72613	1.01784
8	0.61866	1.32865
9	0.47182	2.17626
10	0.33408	3.15374
11	0.20842	4.17075
12	0.06189	5.20418
13	$\bar{1}$.94223	6.24580
14	$\bar{1}$.78913	7.29229
Σ	11.20073	62.54999

Finally, the functions G and F are formed anew from the adjusted values of H, m, q. The adjusted values of s, b and a are computed by equations (40), (41) and (42), as follows:

$$s = [\textstyle\sum_{10}^{14} G_i - 2\sum_{5}^{9} G_i + \sum_{0}^{4} G_i] : [\sum_{10}^{14} F_i - 2\sum_{5}^{9} F_i + \sum_{0}^{4} F_i]$$

$$= [0.33574 - 2 \times 3.72304 + 7.14194] \div [26.06676 - 2 \times 8.62436 + 27.85887]$$

$$= 0.03161 \div 36.67691 = 0.00086185.$$

$$b = [\textstyle\sum_{8}^{14} G_i - \sum_{0}^{6} G_i - s(\sum_{8}^{14} F_i - \sum_{0}^{6} F_i)] : [k(n - k)\Delta t]$$

$$= [1.42623 - 9.04837 - 0.00086185(29.57167 - 31.96048)] \div [7(15 - 7)1]$$

$$= [-7.62214 - 0.00086185 \times (-2.38881)] \div [56] = -0.13607.$$

$$a = [\textstyle\sum_{0}^{14} G_i - b\sum_{0}^{14} t - s\sum_{0}^{14} F_i] : n$$

$$= [11.20073 - (-0.13607 \times 105) - 0.00086185 \times 62.54999] \div 15 = 1.69561.$$

In the present case, the values of H_0, m_0 and q_0 were known within definite limits from previous experimentation. The values of the corrections, $\theta \cdot \Delta$, were, on this account, smaller than should ordinarily be expected from a first application of the technique. Always, it is necessary to take Δ sufficiently large to insure $\theta < 1$. As a preliminary step, it is not infrequently advantageous to compute trial values of ϵ by holding constant each two of the parameters H_0, m_0 and q_0 while experimenting roughly with the third.

TABLE V(A)

Ordinates of Fitted Curves

Year	Census Count	Symmetric Ordinates	Percentage Deviations	Skew Ordinates	Percentage Deviations
1790	3.909	3.88	−0.78	3.87	−0.01
1800	5.280	5.28	−0.03	5.27	−0.25
1810	7.202	7.16	−0.52	7.15	−0.73
1820	9.587	9.69	+1.07	9.67	+0.88
1830	12.866	13.04	+1.37	13.02	+1.20
1840	17.069	17.45	+2.22	17.42	+2.09
1850	23.192	23.15	−0.20	23.13	−0.28
1860	31.443	30.38	−3.36	30.37	−3.42
1870	38.558	39.36	+2.09	39.31	+1.95
1880	50.156	50.18	+0.05	50.07	−0.18
1890	62.948	62.61	−0.31	62.60	−0.55
1900	75.995	76.79	+1.05	76.64	+0.86
1910	92.329	91.76	−0.62	91.72	−0.67
1920	106.001	106.96	+0.90	107.16	+1.09
1930	123.068	121.66	−1.14	122.23	−0.69

TABLE V(B)

Extrapolations

Year	Forecast	Sym. O.	Sk. O.	Year	Sym. O.	Sk. O.
1940	137.20	135.22	136.26	1780	2.844	2.850
1950	149.29	147.18	148.78	1770	2.083	2.095
1960	159.88	157.33	159.52	1760	1.523	1.539
1970	168.71	165.66	168.42	1750	1.113	1.130
1980	175.83	172.33	175.59	1740	0.813	0.829
1990	181.46	177.52	181.25	1730	0.594	0.608
2000	185.82	181.52	185.63	1720	0.434	0.445
2010	189.14	184.55	188.98	1710	0.316	0.280
2020	193.11	186.82	192.97	1700	0.231	0.238
2030	193.54	188.52	193.40	1690	0.168	0.173
2040	194.94	189.77	194.83	1680	0.123	0.127
2050	195.98	190.72	195.87	1620	0.090	0.092
2060	196.75	191.39	196.64	1610	0.065	0.067
2070	197.31	191.88	197.22	1600	0.048	0.049
2080	197.73	192.25	197.64	1590	0.035	0.036
2090	198.03	192.52	197.94			
2100	198.25		198.17			
2110	198.42		198.34			
2120	198.54		198.46			
2130	198.63		198.55			

Part VI

XIII. **General Considerations.** The technique of solution for the numerical values of parameters presented in the foregoing pages is generally applicable to continuous functions of real variables. The abridged procedure may be followed whenever the given function involves a component which is linear in certain of the parameters: for, in such cases, it is always possible to effect a transformation of ordinates which will permit of the elimination of the parameters of the linear component. In any event, the equation of the function may be solved for a single parameter which may then be employed, as in our illustration, as a means of determining the values of the test constant, omega.

XIV. **An Interpretation of Results.** The equations of the symmetric and skew logistic curves as computed for the population of the United States are, written to the natural base, as follows:

$$p = 193.25:[1 + e^{3\ 88826-0\ 31562t}].$$

$$p = 198.80:[1 + e^{3\ 90429-0\ 31331t+0\ 0019845\sqrt{1+1\ 0581^2(t-7\ 1732)^2}}]$$

The amount of skewness in the second of these equations, as measured by the value of s, is small; but, owing to the fair size of the parameter m, it develops rapidly and affects the form of the curve sensibly. The major effect is to raise the value of the limiting population as given in the first equation by about six millions and to prolong the period of growth by about forty years. The approximate limit of 193 millions in the symmetric form is reached about the year 2090; while the approximate limit of 199 millions of the skew form is not arrived at until about the year 2130.

The positive sign of s makes for a decreasing acceleration of the rate of increase during the earlier phases of growth and for an increasing retardation of this rate during the later phases, the value of q fixing the point of transition in the year 1861. This general epoch has often been cited by sociologists as marking the shift from a dominantly rural-agricultural civilization to a dominantly urban-industrial one. The point at which the change takes place has, to my knowledge, never before been defined mathematically.

Both curves fit the observations excellently, as shown by the percentage deviations of Table V(a). The forecasted growth presented in Table V(b) is based on the skew ordinates, the formula being

$$P_t = p_t(P_{14}/p_{14})^{1/(t-14)}, \tag{43}$$

where P denotes the actual population series, observed or predicted, and p, the skew ordinates. The assumptions of the formula are two: first, that it is the observed population P_{14} which initiates the forecasted series; and, second, that the influence of the correction factor P_{14}/p_{14} diminishes with the time.

The extrapolations of both the skew and symmetric formulas contrast with the results obtained by Doctors Dublin and Lotka, who predict a stationary

population of 150 millions by 1970. For the same year, the ordinates of both the skew and symmetric curves exceed this figure, the one by 15.66, and the other by 18.42 millions.

The limit of 150 millions referred to was arrived at by analysis of current tendencies in birth and death rates. The argument is that current birth rates are spuriously high and current death rates spuriously low because of the abnormally high proportion of men and women in the reproductive ages. This circumstance is due, in part, to the influx in the past of immigrants from communities having a high normal birth rate, and, in part, to the high birth rates of preceding generations of parents in this country.

After computation of the necessary corrections has been made, the true rate of natural increase of the white population for the registration area of the United States for the year 1920 is seen to be only about 5.4 per thousand instead of the 10.7 per thousand indicated by the crude rates. For the year 1930, the actual rate of increase is 7.5 per thousand; while the corrected or true rate turns out to be virtually zero. Under the interpretation of the authorities cited, the spurious excess of births over deaths will be entirely dissipated by the year 1970, with the result of the stationary population predicted.

The hazard peculiar to this method of inference arises from two assumptions that are made: first, that the present collection and registration of vital statistics is sufficiently reliable to make precise estimate of the true rate of natural increase possible; second, that the tendencies of fecundity and mortality exhibited by current data are stable.

With respect to the first assumption, the authors have this to say:

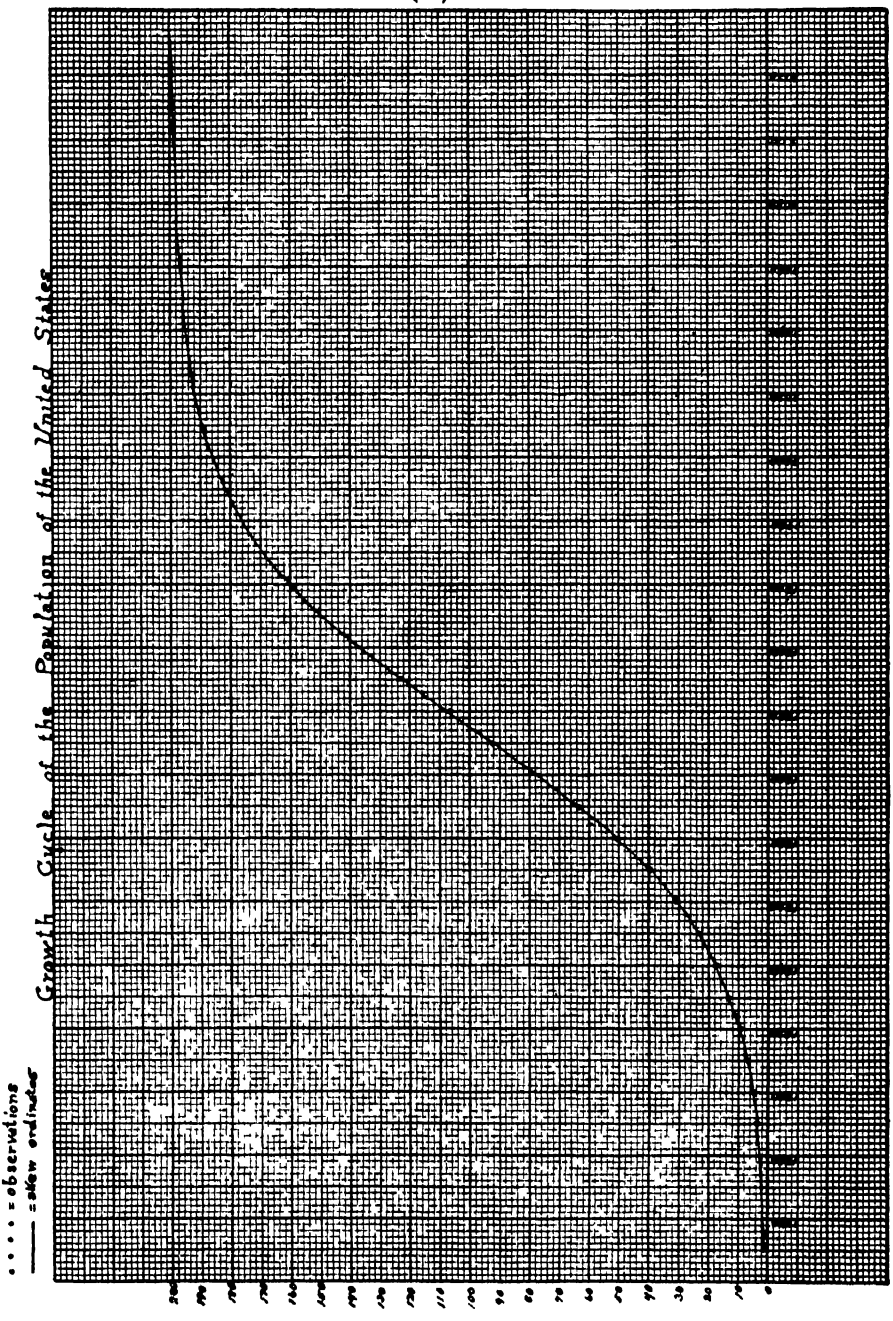
"One factor of safety of unknown magnitude remains. There is still some degree of laxity in the registration of births, and the figures of the true rate of natural increase may, on that account, be somewhat larger than recorded above."

The caution of the authors in this statement is in contrast with the uncritical acceptance of their results by those who fail to grasp the implications of technique.

Concerning the second assumption, it may be pointed out that many of the tendencies exhibited by current data must be regarded as statistically reversible. Falling birth rates due to drift of population to cities, to postponement of marriage on the part of professional classes, to the increasing cost of child culture, to the ubanization of rural life and to the restriction of immigration may be definitely altered by reversals in tendency. The flow of population may move into extraurban and subrural districts, where birth rates are more favorable to increase. The cost of child culture may, in part, be socially assumed. Improvement in economic conditions may lessen the drain on the resources of the family. The tendency for rural birth rates to fall may be checked. Immigration may increase with improving economic conditions. Death rates may be further reduced in many age classes and for many causes.

In fine, when we attempt to project into the future the components that

(24)


Growth Cycle of the Population of the United States
. . . . = observations
—— = ogive ordinates

FIG

determine the trend of natural increase, we encounter risks which vastly exceed those involved in the projection of the population series itself. Most of the data from which component trends must be determined cover but a brief period of time; while population data extends back for a century and a half. In this connection, it is not impertinent to inquire the criterion of relevance that will warrant a rejection of the items of the very series we are seeking to forecast.

It is a cardinal principle of logistic theory that the growth of population depends primarily on the continued supply of basic resources, physical and social, and that the dissipation of these resources is registered in the growth rate of the population itself. Any tendency of a population series toward skewness, that is, toward departure from the symmetric type of growth, is more likely to persist if it is systematic in character. The skew forms of the logistic function which we have developed permit us to measure any existing systematic tendency of the data toward skewness, and, therefore, to improve on the symmetric expectation of future growth.

In the case of the United States population, the evidence of skewness, insofar as it bears on the problem of expectation, is adverse to the conclusion that the ultimate limit of growth will be less than the symmetric asymptote. Conceding the light that the analysis of current tendencies may throw on the probable occurrence of future deviations from trend, the best criterion of long-time growth remains the logistic projection.

This statement, to be sure, does not relieve us of the necessity for recognizing the nature of the hazard that inheres in making a prediction from a trend extrapolation. The hazard involved in this type of inference arises from the assumption that the basic conditions of growth are stable, or, in other words, that the values of the parameters of the forecasting formula will remain substantially unchanged with the inclusion of new observations. Time alone can provide the final test of the continued validity of this assumption.

XV. **The Law of Organic Growth.** The law of organic growth in its most general form may be written:

$$p = L + H{:}[1 + e^{a+bt+s_1u_1+s_2u_2+s_3u_3}], \tag{44}$$

where $u_1 = \sin[m(t + q)]$; $u_2 = \log[1 + m^2(t + q)^2]$; $u_3 = \sqrt{1 + m^2(t + q)^2}$.

For most practical purposes, the evaluation of thirteen parameters is out of the question; hence, the restricted forms α, β, and γ, equation (18), will be the ones most generally employed.

I have made use of the term *law of organic growth* with reference to the logistic forms developed because I believe these functions to be the best means yet devised for the representation of the sequential changes which living organisms regularly manifest as individuals or societies. It states, in a quantitative form, all that is qualitatively implied by the so-called "law of diminishing returns" as this is commonly invoked by economists. The special sense in which I have used the term *law* may be expressed as follows:

A statistical law is a mathematical generalization on the behavior of a system of observations such that the implications of the formula are in accord with the assumptions basic to the phenomenon observed, and such that evaluations of the parameters of the formula determined from random samples are mutually consistent.

A statistical law, then, posits a system of relations manifesting itself in the form of observations which must be subjected to analysis before the true nature of their interrelations can be inferred. It expresses a probable, rather than a certain, inference; but, within the limitations of its claim to precision, it leaves reason no more free to reject its specification of reality than does a law of mechanics. Indeed, the point is still in dispute as to whether any law of science can be more than a statement of probabilities.

In contradistinction, the term *empirical formula* is properly restricted to cover the representation of the single set of observations at hand, and bears no necessary relation to any larger system. A sufficient test of an empirical formula is, therefore, the test of fit.

We may fit an indefinite number of formulas to a population series and obtain satisfactory results so far as agreement is concerned; but, on extrapolating, the same formulas will yield results that are patently absurd. The backward extrapolation for the population of the United States shown in Table V(b) represents the known facts as closely as could be expected when we take into consideration that census enumerations include aboriginal and immigrant populations as well as native born. Certainly, no random empirical formula, selected on the ground of goodness of fit, could be expected to yield as satisfactory a result.

Logistic theory does not, then, profess to guarantee infallibility of prediction. A population is not a mere aggregate of unrelated individuals inhabiting a restricted area, but a unified organization which grows by the utilization of total resources. When the supply of resources is profoundly disturbed or the basis of organizational unity destroyed, then the basis of prediction also is destroyed. And such reasoning is by no means peculiar to the sphere of social organization; for the integrity of any purely mechanical system is likewise conditioned by the assumption that the basis of coherence persists.

At this point, those in whom the speculative disposition is strong may query: if statistical prediction does not yield a certain result, is it, in the final analysis, superior to the ready and far less expensive method of guessing?

In answer, I can only say that, *a posteriori*, we can always, among a sufficiently large batch of guessers, find someone who has guessed well; but how, *a priori*, are we to know the good guesser from the poor? A population series consists of definite magnitudes, and any prediction of its development must result in the selection, out of a vast array of possible magnitudes, that which is most consistent with all the known facts. The gambler may elect to hazard his stake on the result of a random estimate; but the prudent will give heed to the exacting, if laborious, procedure of mathematical analysis.

ADDENDUM

Another solution of the theoretical problem stated in Section I may here be noted.

Given, as before, the function $y = f(x, a, b \cdots)$, we may, by assigning three approximate values to each parameter, compute 3^p sets of values for the function y, thus:

$$y_{11} = f(x, a_1 b_1 \cdots); \qquad y_{12} = f(x, a_1 b_2 \cdots); \qquad y_{13} = (x, a_1 b_3 \cdots); \text{ etc.}$$

From the observations Y, we may compute 3^p sets of the residuals $y - Y$; and from these several sets of residuals, the corresponding standard errors of estimate, σ, may be computed for each set of values of the function y; thus, we have:

$$\sigma_{11} = \phi(Y, x, a_1 b_1)$$

$$\sigma_{12} = \phi(Y, x, a_1 b_2)$$

$$\sigma_{13} = \phi(Y, x, a_1 b_3)$$

Restricting the parameters to a, b, and holding a constant, we observe that the values $\sigma^2_{11}, \sigma^2_{12}, \sigma^2_{13}$ must vary with the assigned values of the parameter b, and take a minimum value when b takes its true or most probable value. As the errors in the approximation to b increase positively and negatively without limit, the computed values of σ^2 will tend toward the infinite. They may, therefore, be assumed to lie on the arc of a parabola whose equation is a quadratic function of xa_1b; hence, we may form the following equations of representation:

$$\sigma^2_{11} = k_{11} + l_{11}a_1 + m_{11}a_1^2.$$

$$\sigma^2_{12} = k_{12} + l_{12}a_1 + m_{12}a_1^2.$$

$$\sigma^2_{13} = k_{13} + l_{13}a_1 + m_{13}a_1^2.$$

By addition, we have,

$$\sigma^2_{11} + \sigma^2_{12} + \sigma^2_{13} = k_{11} + k_{12} + k_{13} + (l_{11} + l_{12} + l_{13})a_1 + (m_{11} + m_{12} + m_{13})a_1^2.$$

By appropriate variations in subscript, similar equations may be written in a_2 and a_3, thus:

$$\sigma^2_{21} + \sigma^2_{22} + \sigma^2_{23} = k_{21} + k_{22} + k_{23} + (l_{21} + l_{22} + l_{23})a_2 + (m_{21} + m_{22} + m_{23})a_2^2.$$

$$\sigma^2_{31} + \sigma^2_{32} + \sigma^2_{33} = k_{31} + k_{32} + k_{33} + (l_{31} + l_{32} + l_{33})a_3 + (m_{31} + m_{32} + m_{33})a_3^2.$$

These three equations are all of the quadratic form, and may be conveniently written as follows:

$$A_1 = K_1 + L_1a_1 + M_1a_1^2.$$

$$A_2 = K_1 + L_1a_2 + M_1a_2^2.$$

$$A_3 = K_1 + L_1a_3 + M_1a_3^2.$$

By precisely similar reasoning, the following equations in b may be developed:

$$B_1 = K_2 + L_2 b_1 + M_2 b_1^2.$$

$$B_2 = K_2 + L_2 b_2 + M_2 b_2^2.$$

$$B_3 = K_2 + L_2 b_3 + M_2 b_3^2,$$

where

$$B_1 = \sigma_{11}^2 + \sigma_{21}^2 + \sigma_{31}^2; \quad B_2 = \sigma_{12}^2 + \sigma_{22}^2 + \sigma_{32}^2; \quad B_3 = \sigma_{13}^2 + \sigma_{23}^2 + \sigma_{33}^2.$$

Since the values of a_1, a_2, a_3 and b_1, b_2, b_3 are assigned, the two sets of equations may each be simultaneously solved to obtain values for K_1, L_1, M_1 and K_2, L_2, M_2. To obtain the conditions for A = a minimum, B = a minimum, we differentiate with respect to a and b, as follows:

$$D_a(A) = L_1 + 2M_1 a; \qquad D_b(B) = L_2 + 2M_2 b.$$

Setting these two equations equal to zero and solving, we obtain the adjusted values of a and b, thus:

$$a = -L_1 : 2M_1; \qquad b = -L_2 : 2M_2.$$

The extension of this method to the case of p parameters is obvious. Assigning three approximations to each parameter, we hold constant a value of one parameter (say a_1), we form all possible combinations of subscripts for the remaining parameters ($b_1 b_2 b_3$ with $c_1 c_2 c_3$ with etc.). This will yield 3^{p-1} values of σ^2, each of which is associated with a_1. Repeating this process, we can form similar sets of values of σ^2 by association with a_2 and a_3. We can then form the sums $A_1 = \sigma(Yxa_1bc \cdots)$; $A_2 = \sigma(Yxa_2bc \cdots)$; $A_3 = \sigma(Yxa_3bc \cdots)$. In all, $3 \times 3^{p-1}$ or 3^p distinct determinations of σ^2 will be required. In like manner, the equations for B_1, B_2, B_3 and C_1, C_2, C_3, etc. are formed. The solutions for the adjusted values $a, b, c, \cdots$ follow directly.

Since the method of solution given in Part I requires the computation of but 2^p values of σ^2, it is evident that the method of this section is the more onerous when considering the determination of a single set of adjusted values of parameters, the excess being of the order $3^p : 2^p = (1.5)^p$. However, being more precise, the present method will require fewer approximations to arrive at satisfactory values of the parameters sought. In other words, the mathematical advantage of economy lies with the theta technique; while the advantage of precision lies with the quadratic technique.

SELECTED BIBLIOGRAPHY

DUBLIN, L. I. AND LOTKA, A. J. On the True Rate of Natural Increase. J. A. S A. S. 1925.

The True Rate of Natural Increase of the Population of the United States Metron, Je. 1930.

HOTELLING, H. Differential Equations Subject to Error and Population Estimates. J. A. S. A. 1927.

HOTELLING, H. AND F. Causes of Birth Rate Fluctuations. J. A. S. A. 1931.
KNIBBS, G. H. Laws of Growth of a Population. J. A. S. A. 1926-27.
LEHFELDT, R. A. The Normal Law of Progress; J. R. S. S. 1916.
LOTKA, A. J. Studies on the Mode of Growth of Material Aggregates. A. J. Sci. 1907.
PEARL, R. AND REED, L. J. On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation. Proc. N. A. Sci. 1920.
On the Mathematical Theory of Population Growth. Metron, 1923.
WILL, H. S. On Fitting Curves to Observational Series by the Method of Differences; Ann. M. S. My. 1930.
WOLFE, A. B. Is there a Biological Law of Human Population Growth? Q. J. Ec. Ag. 1927.
YULE, G. U. The Growth of Population and the Factors which Control It. J. R. S. S. Jn. 1924.

ON A METHOD FOR EVALUATING THE MOMENTS OF A BERNOULLI DISTRIBUTION[1]

By Everett H. Larguier, S.J.

1. The moments (per unit frequency) of a frequency distribution have long been regarded as useful characteristics of the distribution. If we denote the moment about the arithmetic mean by μ, we have for the Bernoulli distribution

$$\mu_s = \sum_{x=0}^{n} (\bar{x})^s f(x),$$

where $\bar{x} = x - np$ and $f(x) = \binom{n}{x} p^x q^{n-x}$.

To evaluate the s-th moment about the arithmetic mean has always been a laborious task. Karl Pearson[2] gave the s-th moment about the arithmetic mean as,

$$\mu_s = \left[\frac{d^s}{dx^s}[qe^{px} + pe^{-qx}]^n\right]_{x=0}, \tag{1}$$

which he said at that time was perhaps the easiest expression for obtaining these moment coefficients by successive differentiation. Romanovsky,[3] however, was able to develop the recursion formula,

$$\mu_{s+1} = pq\left[ns\mu_{s-1} + \frac{d\mu_s}{dp}\right], \tag{2}$$

for the moments about the mean. Another relation for these moments is

$$\mu_{s+1} = \sum_{i=0}^{s-1} \binom{s}{i} [npq\mu_i - p\mu_{i+1}]. \tag{3}$$

Recently Kirkham[4] gave the expressions for the first eight moments which, however, are not in a form well adapted for numerical calculation on a machine.

[1] Presented to the American Mathematical Society, January 2, 1936.

[2] Karl Pearson, *Biometrika*, vol. 12 (1918–1919), footnote, p. 270. This expression is obtained from the moment-generating function. Obviously this method is exceedingly impractical for numerical calculations.

[3] V. Romanovsky, "Note on the moments of the binomial $(p + q)^n$ about its mean," *Biometrika*, vol. 15 (1923). Recently this expression was given a simple proof by A. T. Craig (*Bull. Amer. Math. Soc.*, vol. 40, pp. 262–264) and extended to the Poisson case.

[4] W. J. Kirkham, "Moments about the arithmetic mean of a binomial frequency distribution," *Annals of Mathematical Statistics*, vol. VI, pp. 96–101.

2. It is the purpose of this paper to express the s-th moment about the arithmetic mean in the form

$$\mu_s = \sum_{t=1}^{t=s} F_{s,t}(n)p^t, \tag{4}$$

where $F_{s,t}(n)$ are determinable functions of n dependent on s and t. We note here that p and q are the probabilities of the success and failure of an event in a single trial.

Since we know that $\mu_2 = npq$ and $\mu_1 = 0$, it is evident that the part of (2) enclosed in [ ] will be of degree 2 less than $s + 1$ in p and hence (4) will satisfy as a representation of the moment.

3. To obtain a recursion formula for the functions $F_{s,t}(n)$ we differentiate (4) with respect to p. This gives

$$\frac{d\mu_s}{dp} = \sum_{t=1}^{s} tF_{s,t}(n)p^{t-1}.$$

By (2) we may then write

$$\begin{aligned}\sum_{t=1}^{s+1} F_{s+1,t}(n)p^t &= p(1-p)ns\sum_{t=1}^{s-1} F_{s-1,t}(n)p^t + p(1-p)\sum_{t=1}^{s} tF_{s,t}(n)p^{t-1} \\ &= ns\sum_{t=2}^{s} F_{s-1,t-1}(n)p^t - ns\sum_{t=3}^{s+1} F_{s-1,t-2}(n)p^t \\ &\quad + \sum_{t=1}^{s} tF_{s,t}(n)p^t - \sum_{t=2}^{s+1}(t-1)F_{s,t-1}(n)p^t.\end{aligned}$$

Since this is an identity in p, we have immediately the following recursion formula for determining $F_{s,t}(n)$:

$$\begin{aligned}F_{s,t}(n) &= n(s-1)F_{s-2,t-1}(n) - n(s-1)F_{s-2,t-2}(n) + tF_{s-1,t}(n) \\ &\quad - (t-1)F_{s-1,t-1}(n)\end{aligned} \tag{5}$$

in which

$$F_{0,0}(n) = 1; \quad \text{and} \quad F_{s,t}(n) = 0 \quad \text{for} \quad \begin{cases} t > s; \\ t < 1,\ s > 0; \\ t = 1,\ s = 1. \end{cases} \tag{6}$$

These definitions arise from the known values of the moments and the conditions imposed by the identity in p.

By means of (5) and (6) we are able to obtain very readily the values for $F_{s,t}(n)$ which are given in Table 1.

TABLE I

Values of $F_{s,t}(n)$

s	$F_{s,1}(n)$	$F_{s,2}(n)$	$F_{s,3}(n)$	$F_{s,4}(n)$
1	0	0	0	0
2	n	$-n$	0	0
3	n	$-3n$	$2n$	0
4	n	$-7n + 3n^2$	$12n - 6n^2$	$-6n + 3n^2$
5	n	$-15n + 10n^2$	$50n - 40n^2$	$-60n + 50n^2$
6	n	$-31n + 25n^2$	$180n - 180n^2 + 15n^3$	$-390n + 415n^2 - 45n^3$
7	n	$-63n + 56n^2$	$602n - 686n^2 + 105n^3$	$-2100n + 2590n^2 - 525n^3$
8	n	$-127n + 119n^2$	$1932n - 2394n^2 + 490n^3$	$-10206n + 13895n^2 - 3850n^3 + 105n^4$

s	$F_{s,5}(n)$	$F_{s,6}(n)$
1	0	0
2	0	0
3	0	0
4	0	0
5	$24n - 20n^2$	0
6	$360n - 390n^2 + 45n^3$	$-120n + 130n^2 - 15n^3$
7	$3360n - 4270n^2 + 945n^3$	$-2520n + 3234n^2 - 735n^3$
8	$25200n - 35700n^2 + 10990n^3 - 420n^4$	$-31920n + 46004n^2 - 14770n^3 + 630n^4$

s	$F_{s,7}(n)$	$F_{s,8}(n)$
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	$720n - 924n^2 + 210n^3$	0
8	$20160n - 29232n^2 + 9520n^3 - 420n^4$	$-5040n + 7308n^2 - 2380n^3 + 105n^4$

With this table it is a relatively easy task to evaluate the first eight moments with the aid of a calculating machine.

4. As an illustration of the preceding we propose to evaluate the first eight moments about the arithmetic mean for the binomial, $(.06785 + .93215)^{378}$. We first evaluate the coefficients $F_{s,t}(n)$.

TABLE II[5]

Values of $F_{s,t}(378)$

s	$F_{s,1}(378)$	$F_{s,2}(378)$	$F_{s,3}(378)$	$F_{s,4}(378)$	$F_{s,5}(378)$
1	0	0	0	0	0
2	378	−378	0	0	0
3	378	−1,134	756	0	0
4	378	426,006	−852,768	426,384	0
5	378	1,423,170	−5,696,460	7,121,520	−2,848,608
6	378	3,560,382	784,501,200	−2,371,307,400	2,374,868,160
7	378	7,977,690	5,573,275,090	−27,986,054,000	50,430,749,000
8	378	16,955,190	26,123,640,500	1,937,705,370,000	−7,986,171,610,000

s	$F_{s,6}(378)$	$F_{s,7}(378)$	$F_{s,8}(378)$
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	−791,622,720	0	0
7	−39,236,327,400	11,210,379,300	0
8	12,070,808,800,000	−8,064,644,270,000	2,016,161,070,000

Then running off the powers of p, we have:

$p = .067\ 85$ $\qquad$ $p^5 = .000\ 001\ 437\ 968\ 13$

$p^2 = .004\ 603\ 622\ 5$ $\qquad$ $p^6 = .000\ 000\ 097\ 566\ 137\ 6$

$p^3 = .000\ 312\ 355\ 787$ $\qquad$ $p^7 = .000\ 000\ 006\ 619\ 862\ 44$

$p^4 = .000\ 021\ 193\ 340\ 1$ $\qquad$ $p^8 = .000\ 000\ 000\ 449\ 157\ 667$

Applying (4) we have

[5] In this table, as well as in the one that follows, all values are correct to nine significant figures.

TABLE III

Values of $p^t F_{s,t}(378)$

s	2	3	4	5
$pF_{s,1}(378)$	25.6473	25.6473	25.6473	25.6473
$p^2F_{s,2}(378)$	−1.7401693	−5.2205079	1961.17087	6551.73743
$p^3F_{s,3}(378)$	0.	.2361410	−266.36702	−1779.32225
$p^4F_{s,4}(378)$	0.	0.	9.03650	150.92880
$p^5F_{s,5}(378)$	0.	0.	0.	−4.09621
$p^6F_{s,6}(378)$	0.	0.	0.	0.
$p^7F_{s,7}(378)$	0.	0.	0.	0.
$p^8F_{s,8}(378)$	0.	0.	0.	0.
μ_s	23.9071307	20.6629331	1729.48765	4944.89507

s	6	7	8
$pF_{s,1}(378)$	25.647	25.65	25.6
$p^2F_{s,2}(378)$	16390.655	36726.27	78055.3
$p^3F_{s,3}(378)$	245043.490	1740844.73	8159870.3
$p^4F_{s,4}(378)$	−50255.924	−593117.96	41066448.9
$p^5F_{s,5}(378)$	3414.985	72517.81	−11483860.3
$p^6F_{s,6}(378)$	−77.236	−3828.14	1177702 2
$p^7F_{s,7}(378)$	0.	74.21	−53386.8
$p^8F_{s,8}(378)$	0.	0.	905.6
μ_s	214541.617	1253242.57	38945760.8

This gives us the desired moments about the arithmetic mean of the binomial $(.06785 + .93215)^{378}$. These values may be rapidly checked by applying (3) to μ_8.

SAINT LOUIS UNIVERSITY,
SAINT LOUIS, MISSOURI.

A METHOD OF DETERMINING THE REGRESSION CURVE WHEN THE MARGINAL DISTRIBUTION IS OF THE NORMAL LOGARITHMIC TYPE

By Carl-Erik Quensel

Assistant at the Statistical Institute of the University of Lund, Sweden

In a paper[1] in this Journal Professor S. D. Wicksell gave the general outlines of a new method of calculating the regression lines. This problem was later on treated in detail by Dr. Walter Andersson.[2] His method was to develop the formulas for the regression lines into a series of orthogonal polynomials under the assumption that the marginal distribution of the independent variate belonged to certain mathematically defined distributions, and to determine the constants with the aid of the method of the least squares.

Among other cases he treated also the case where the marginal distribution was of the normal logarithmic type:

$$F(x) = \frac{\log e}{\sigma_l \sqrt{2\pi}\,(x-a)}\, e^{-\frac{1}{2}\left[\frac{\log(x-a)-l}{\sigma_l}\right]^2}. \tag{1}$$

But as his method is entirely different from the method I shall give here, I will not go any further into the method used by Dr. Andersson.

When the correlation surface $F(x, y)$ of the variates x and y is given and then of course also the marginal distribution of x, $F(x)$, it is known that the mean y_x of the dependent variate y in an infinitely small array with the value of x between x and $x + dx$ is given as a function of the independent variate x by the following formula (2)

$$y_x = \frac{\int y F(x, y)\, dy}{\int F(x, y)\, dy}. \tag{2}$$

In this formula the integrals are to be extended over the whole domain of the variation of y.

If now we make any transformation of x by introducing a new variate u, related to x by the formula $u = \psi(x)$, where we must suppose that u is a one-valued function of x and contrary, the distribution $f(u, y)$ of the variate u and y is given by the relation

$$f(u, y)\, du\, dy = F(x, y)\, dx\, dy \tag{3}$$

[1] S. D. Wicksell. Remarks on Regression. Annals of Mathematical Statistics, 1930.

[2] Walter Andersson. Researches into the theory of Regression. Meddelande från Lunds Astronomiska Observatorium. Ser II. N:r 64.

Writing the formula (2) in the following form:

$$y_x \quad \frac{\int yF(x, y)\,dx\,dy}{\int F(x, y)\,dx\,dy};$$

we see at once that the mean y_x can be given as the following function of u:

$$y_x = \frac{\int yf(u, y)\,dy}{\int f(u, y)\,dy} \tag{4}$$

This relation, of course, is self-evident. The mean of the dependent variate in an array of the independent variate will be unchanged, when we change the variate x for another variate u, related to x by a one-valued function.

The problem of finding the regression line of the mean y_x can in such a way be much simplified, if it is possible to make a favorable transformation of the independent variate x.

As shown by Professor Wicksell[3] we may, under certain conditions concerning the marginal distribution $f(u)$, write the expression of the regression line in the following form:

$$y_x = \sum_{0}^{\infty} (-1)^n \frac{\lambda_{n,1}}{n!} \frac{f^{(n)}(u)}{f(u)} \tag{5}$$

where the $\lambda_{n\,1}$ coefficients are the seminvariants of the distribution of u and y.

The conditions which the function $f(u)$ must satisfy are among others that the function and all its derivates are continuous in the domain of variation and that the function and its derivates disappear in the limits of that domain. These conditions are satisfied by the normal curve of error.

In the case where the distribution of u is normal, the derivates $f^{(n)}(u)$ take the following form:

$$f^{(n)}(u) = (-1)^n H_n(u) f(u); \tag{6}$$

where the expressions $H_n(u)$ are the well known Hermitian polynomials.

The formula (5) takes the following simple form.

$$y_x = \sum_{0}^{\infty} \frac{\lambda_{n,1}}{n!} H_n(u) \tag{7}$$

If we can change the given marginal distribution $F(x)$ by a favorable substitution $u = \psi(x)$ into a normal curve, and if, this substitution made, we can

[3] S. D. Wicksell. Analytical Theory of Regression. Meddelande från Lunds Astronomiska Observatorium. Ser II. N:r 69.

calculate the coefficients $\lambda_{n,1}$ from the moments or other known characteristics of the given correlation distribution, $F(x, y)$, it is possible to express the regression line as the formula (8) shows:

$$y_x = \sum_0^\infty \frac{\lambda_{n,1}}{n!} H_n[\psi(x)] \tag{8}$$

It must be observed that the polynomials $H_n[\psi(x)]$ are orthogonal with regard to the distribution $F(x)$ of the independent variate x. We have

$$\int H_i[\psi(x)]\, H_j[\psi(x)]\, F(x)\, dx = \int H_i(u)\, H_j(u) f(u)\, du = 0 \qquad i \neq j$$

Not in all cases it will perhaps be possible to calculate the $\lambda_{n,1}$ coefficients, when we have transformed the marginal distribution into the normal curve, but in one case it is rather simple to calculate these coefficients from the moments given.

The case alluded to is the one, where the variate u is given from x by the relation $u = \log(x - a)$, that is that the marginal distribution is of the so called normal logarithmic type (1).

In that case it is possible to calculate the $\lambda_{n,1}$ coefficients from the marginal moments $V_{n,0}$ and from the correlation moments of the type $V_{n,1}$.

We suppose that the marginal distribution is of the logarithmic type and that from the moments of the x distribution we have determined the three constants a, σ_l and l in the usual manner.[4]

Then we calculate from the given correlation distribution the moments $V'_{n,0}$ about the point $x = a$ and the correlation moments $V'_{n,1}$ about the point $x = a$ and $y = m_y$ (the mean value of the y-variate).

From these moments it is possible to calculate the $\lambda_{n,1}$ coefficients in the following way.

The characteristic function of u and y is given by the following relation:

$$U(t_1 t_2) = e^{\Sigma \frac{\lambda_{kl}}{k!\,l!}\, t_1^k\, t_2^l} = \int\int e^{t_1 u + t_2 y} f(u, y)\, du\, dy \tag{9}$$

where the integrals are extended over the whole domain of variation.

If the distribution of u is according to the normal law, we have $\lambda_{k,0} = 0$ for $k \geqq 3$, but in the calculations here it is not at all necessary to suppose anything about these higher seminvariants. On the other side, the correlation distribution $f(u, y)$ is obtained from the characteristic function by the inversion theorem.

$$f(u, y) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\Sigma \frac{\lambda_{kl}}{k!\,l!}\, (i w_1)^k (i w_2)^l}\, e^{-i w_1 u - i w_2 y}\, dw_1\, dw_2 \tag{10}$$

[4] How these are to be determined is shown in Pae- Tsi Yuan. On the logarithmic Frequency distribution and the Semi-Logarithmic Correlation Surface. Annals of Mathematical Statistics, 1933.

But we can also get the following relation

$$\int e^{it_1 u} f(u, y)\, du = \frac{1}{2\pi} \int e^{-iw_2 y} e^{\sum \frac{\lambda_{kl}}{k!\,l!} t_1^k (i w_2)^l}\, dw_2 \tag{11}$$

Of this last expression (11) between the characteristic function and the distribution function I will make use in the following.

The moments $V'_{s,t}$ of the distribution $F(x, y)$ about the point $x = a$, $y = m_y$ are given by the formula

$$V'_{s,t} = \int\int (x - a)^s (y - m_y)^t F(x, y)\, dx\, dy. \tag{12}$$

If we write y instead of $y - m_y$ and instead of $x - a$ we write e^{bu} ($b = {}_{10}\log e$) the expression (12) takes the following form:

$$V'_{s,t} = \int\int e^{sbu} y^t f(u, y)\, du\, dy \tag{13}$$

For the marginal moments of x about the point $x = a$ we get

$$V'_{n,0} = \int_a^\infty (x - a)^n F(x)\, dx = \int_{-\infty}^\infty e^{nbu} f(u)\, du \tag{14}$$

Comparing this formula (14) with the expression for the characteristic function of the distribution $f(u)$

$$U(t_1) = \int_{-\infty}^\infty e^{it_1 u} f(u)\, du = e^{\sum \frac{\lambda_{k,0}}{k!} t_1^k}; \tag{15}$$

we find the following simple relation

$$V'_{n,0} = e^{\sum \frac{\lambda_{k,0}}{k!} (nb)^k} \tag{16}$$

For the moments of the type $V'_{n,1}$ we get

$$V'_{n,1} = \int\int e^{nbu} y f(u, y)\, du\, dy = \int y dy \int e^{nbu} f(u, y)\, du. \tag{17}$$

If we compare the last integral in the formula (17) $\int e^{nbu} f(u, y) du$ with the formula (11) we see that we can write (17) as follows:

$$V'_{n,1} = \frac{1}{2\pi} \int y dy \int e^{-iw_2 y} e^{\sum \frac{\lambda_{kl}}{k!\,l!} (nb)^k (i w_2)^l}\, dw_2 \tag{18}$$

From the sum $\sum \frac{\lambda_{kl}}{k!\,l!} (nb)^k (iw_2)^l$ we may take out the part $\sum \frac{\lambda_{k,0}}{k!} (nb)^k$,

where l is zero and which therefore does not contain any dignity of w_2, and write the remainder in the following form:

$$\sum \frac{\lambda_l'}{l!} (iw_2)^l$$

where we have

$$\lambda_1' = \lambda_{11} nb + \frac{\lambda_{21}}{2!} (nb)^2 + \frac{\lambda_{31}}{3!} (nb)^3 \cdots$$

$$\frac{\lambda_2'}{2!} = \frac{\lambda_{02}}{2!} + \frac{3\lambda_{12}}{3!} nb + \frac{6\lambda_{22}}{4!} (nb)^2 \cdots$$

The integral $\frac{1}{2\pi} \int e^{-iw_2 y} e^{\sum \frac{\lambda_l'}{l!} (i w_2)^l}$ may be considered as a frequency distribution $\varphi(y)$ with the seminvariants λ_l'.

The formula (18) will thus be written

$$V_{n,1}' = e^{\sum \frac{\lambda_{k0}}{k!} (nb)^k} \int y dy \, \varphi(y) \tag{19}$$

According to (16) we have

$$e^{\sum \frac{\lambda_{k0}}{k!} (nb)^k} = V_{n,0}'$$

and as

$$\int y dy \, \varphi(y) = \lambda_1' = \lambda_{11} nb + \frac{\lambda_{21}}{2!} (nb)^2 + \frac{\lambda_{31}}{3!} (nb)^3 \cdots$$

we get

$$V_{n,1}' = V_{n,0}' \cdot \lambda_1' \tag{20}$$

or

$$\frac{V_{n,1}'}{V_{n,0}'} = \lambda_{11} nb + \frac{\lambda_{21}}{2!} (nb)^2 + \frac{\lambda_{31}}{3!} (nb)^3 \cdots \tag{21}$$

We see that in the formulas for $V_{n,1}'$ we have all the seminvariants $\lambda_{n,1}$ involved. A successive determination of the seminvariants $\lambda_{n,1}$ with the aid of the moments of the same and lower degree is therefore not possible.

However, when we use the formula (8) for the regression, we must suppose that the seminvariants $\lambda_{n,1}$ with growing n converge rather soon towards zero.

If the successive differences $\Delta^n \left(\frac{V_{n,1}'}{V_{n,0}'} \right)$ of the quotients $\frac{V_{n,1}'}{V_{n,0}'}$ are calculated, it may be possible to judge, how far it is possible to go with success. These differences will in most cases diminish rather soon and we shall therefore in most cases get a value of n about which we can suppose that the differences of higher order than this will all be so small that they can be neglected and as a consequence of this fact all higher seminvariants can be neglected too.

When this value of n has been determined, the n first seminvariants will all be obtained from the n first quotients $\frac{V'_{n,1}}{V'_{n,0}}$.

Thus we finally get the regression line as follows:

$$y_x = m_y + \sum_1^n \frac{\lambda_{i,1}}{i!} H_i[\log (x - a) - l]$$

or in standardized units:

$$y_x = m_y + \sum_1^n \frac{\lambda_{i,1}}{i!\,\sigma_l^i} H_i\left[\frac{\log (x - a) - l}{\sigma_l}\right]$$

THE STANDARD ERROR OF A "SOCIAL FORCE"

By Stuart C. Dodd

I. Definitions

In the theory of measurement of social forces certain special cases of frequent occurrence where the population shifts from one date of measurement to the next require the derivation of appropriate standard error formulae.

The theory may be briefly restated[1] in equations as follows: any measurable social change, C, in a population, P, may be defined as the difference in mean scores, S, from surveys or measurements on the dates denoted by subscripts

$$C_{2-1} = S_2 - S_1 = \frac{\Sigma s_2}{P} - \frac{\Sigma s_1}{P} \tag{1}$$

The momentum of a social change may be defined as the product of its time rate in years and the population that is being changed

$$M_{2-1} = PV_{2-1} \tag{2}$$

$$= \frac{PC_{2-1}}{Y_{2-1}} = \frac{P}{Y_{2-1}}(S_2 - S_1) \tag{2a}$$

where Y_{2-1} is the period from date 1 to date 2 and V is the velocity, or speed of change, in that period. The acceleration of a social change is definable as the rate of change of the velocity of change

$$A = \frac{V_{4-3} - V_{2-1}}{.5Y_{(4-3-2+1)}} \tag{3}$$

where each velocity, being an average for its period, is taken as representing the mid-date of that period.

The resultant social force which produces a measured change is now definable as that which accelerates the change in a population. It is measurable as the product of the acceleration and the population.[2]

$$F = AP \tag{4}$$

$$= \frac{P}{.5Y_{(4-3-2+1)}}\left(\frac{S_1}{Y_{2-1}} - \frac{S_2}{Y_{2-1}} - \frac{S_3}{Y_{4-3}} + \frac{S_4}{Y_{4-3}}\right) \tag{5}$$

---

[1] *A Controlled Experiment on Rural Hygiene in Syria*, Dodd, S. C., Publications of the American University of Beirut, Syria, Social Science Series No. 7, 1934, pp. 336.

Also, *A Theory for the Measurement of Some Social Forces*, Dodd, S. C., Scientific Monthly, Vol. XLIII; No. 1, July 1936, pp. 58–62

[2] Force thus defined in terms of its effect is a resultant force, i.e., the residual force after deducting all resisting forces from the total force in the direction of the change observed. This formula defines quantitatively and exactly the "net" force not the "gross" force

II. The Sampling error of one case (momentum)

The formulae for the standard errors of sampling for the above concepts, social change, velocity, momentum, acceleration and force, (C, V, M, A, and F) have been published for the case where the population, P, is the same on all dates of measurement. But it is not always possible to observe the ideal experimental technic of holding the population unchanged in number nor to select out individuals common to all the surveys and to neglect the rest. Ordinarily there will be different P's, P_1, P_2, P_3, and P_4, at the different dates.

To derive the standard errors of (2) and (4) when P shifts, each P is considered to be a sub-sample[3] of the main sample which is $(P_1 + P_2 + P_3 + P_4)$. The orthodox view of sampling is taken where the sub-samples may differ in size but maintain fixed proportions in each main sample which is drawn from the "parent" population.

Let primes denote an M, or other function of (1) to (5), which is an approximation due to the shifting of the population and the use of an average P.

To simplify and generalize the notation, let k denote the constant term compounded of P's and Y's which is associated with each S. The first subscript of k denotes the function, f, which is any particular one of the left hand members of equations (1) to (5) and the second subscript denotes the date of its S. Thus, from (2a)

$$k_{M1} = \frac{-P_1 + P_2}{2Y_{2-1}} = -k_{M2} \tag{6}$$

Then (2) may be rewritten:

$$M'_{2-1} = S_1 k_{M1} + S_2 k_{M2} \tag{7}$$

$$= \sum_1^2 S k_M. \tag{7a}$$

To derive the standard error of (7) the total differential is:

$$dM'_{2-1} = k_{M1}\, d\left(\frac{\Sigma s_1}{P_1}\right) + k_{M2}\, d\left(\frac{\Sigma s_2}{P_2}\right) \tag{8}$$

If Q_{12} denotes the population common to both dates of measurement so that:

$$\begin{aligned} P_1 &= Q_{12} + Q_1 \\ P_2 &= Q_{12} + Q_2 \end{aligned} \tag{9}$$

producing the change. It thus measures only the *observable part* of the total forces in the situation. The fundamental problem remains, as always in science, to observe more adequately, to devise experimental and statistical technics for measuring the different forces (in isolation and in combinations) which facilitate or resist the measured change.

[3] The author is indebted to Mr. S. S. Wilks (Princeton) for this method of deriving these standard errors in a fluctuating population.

and, since the differential of a sum is the sum of the differentials of the several terms, (8) becomes

$$dM'_{2-1} = \frac{k_{M1}}{P_1}\left(\sum_1^{Q_{12}} ds_1 + \sum_1^{Q_1} ds_1\right) + \frac{k_{M2}}{P_2}\left(\sum_1^{Q_{11}} ds_2 + \sum_1^{Q_2} ds_2\right) \tag{10}$$

Squaring gives

$$(dM'_{2-1})^2 = \frac{k_{M1}^2}{P_1^2}(\sum ds_1)^2 + \frac{k_{M2}^2}{P_2^2}(\sum ds_2)^2 + \frac{2\,k_{M1}\,k_{M2}}{P_1\,P_2}\left[\sum_1^{Q_{12}} ds_1 \sum_1^{Q_{12}} ds_2 + \sum_1^{Q_{12}} ds_1 \sum_1^{Q_2} ds_2 + \sum_1^{Q_1} ds_1 \sum_1^{Q_{12}} ds_2 + \sum_1^{Q_1} ds_1 \sum_1^{Q_2} ds_2\right] \tag{11}$$

On summing and dividing by the number of cases to get the expected values, the last three terms in the square brackets vanish. Using the relation where, in random sampling, the correlation between two variables is the same as the correlation between their means

$$r_{12} = r_{S_1S_2} = \frac{\Sigma\, S_1S_2}{Q_{12}\,\sigma_1\sigma_2} = \frac{\Sigma\left(\frac{\Sigma\, s_1}{Q_{12}} \cdot \frac{\Sigma\, s_2}{Q_{12}}\right)}{Q_{12}\,\frac{\sigma_1\,\sigma_2}{\sqrt{Q_{12}\cdot Q_{12}}}} \tag{12}$$

gives

$$\sigma^2_{M'_{2-1}} = \frac{k_{M1}^2\,\sigma_1^2}{P_1} + \frac{k_{M2}^2\,\sigma_2^2}{P_2} + \frac{2\,k_{M1}\,k_{M2}\,Q_{12}\,\sigma_1\,\sigma_2\,r_{12}}{P_1\,P_2} \tag{13}$$

Standard error of momentum when the population shifts

The best estimates of σ_1 and σ_2 are the standard deviations of the scores, s_1 and s_2, and the best estimate of r_{12} is, strictly, the covariance of the common cases divided by the two sigmas. Unless the selection of Q_{12} out of P_1 and P_2 curtails the range in some way (i.e., Q_{12} is not a random selection), then, except for sampling variation, σ_1 and σ_2 are the same in the Q_{12} population as in the P_1 and P_2 populations so that there is only a sampling discrepancy between the ratio above and the r_{12}, the observed correlation between the s_1 and s_2 scores in the Q_{12} population.

III. The generalized standard error

The above standard error may be readily generalized. Any of the equations (1) to (5) may be expressed as a simple linear sum of the products of a variable, S, and its appropriate constant, k.

$$f = \sum_{i=1}^{i=n} S_i\,k_{fi} \tag{14}$$

where f is any one of the concepts S, C, V, A, M or F defined by (1) to (5) and n is the number of surveys, or different S's involved, and i denotes each survey in turn from 1 to n. Thus where f means F, (5) becomes:

$$f'_F = F' = k_{F1} S_1 + k_{F2} S_2 + k_{F3} S_3 + k_{F4} S_4 = \sum_{i=1}^{i=4} k_{Fi} S_i \tag{15}$$

where

$$k_{F1} = -k_{F2} = \frac{P_1 + P_2 + P_3 + P_4}{2\, Y_{(4-3-2+1)}\, Y_{(2-1)}} \tag{16a}$$

$$k_{F4} = -k_{F3} = \frac{P_1 + P_2 + P_3 + P_4}{2\, Y_{(4-3-2+1)}\, Y_{(4-3)}}. \tag{16b}$$

In the special case when a force, F, has been determined from only three surveys using two consecutive periods, $n = 3$ and

$$k_{1F} = \frac{P_1 + P_2 + P_3}{1.5\, Y_{(3-1)}\, (Y_{2-1})} \tag{16c}$$

$$k_{F2} = -\frac{(P_1 + P_2 + P_3)\,(Y_{(2-1)} + Y_{(3-2)})}{1.5\, Y_{(3-1)}\, Y_{(3-2)}\, Y_{(2-1)}} \tag{16d}$$

$$k_{F3} = \frac{P_1 + P_2 + P_3}{1.5\, Y_{(3-1)}\, Y_{(3-2)}} \tag{16e}$$

If the difference between two forces (or other functions, f) has been measured in either the same or in different populations and the significance of the difference in terms of its standard error is desired, f of (14) can also denote that difference.

$$f_{dF} = F_a - F_b; \qquad f_{dM} = M_a - M_b; \text{ etc.} \tag{17}$$

It is only necessary to write the difference as a linear sum of products of S and k on the model of (2a) or (5) to get the k-values for that particular f.

It is now possible to write the standard error formula for f in a single generalized form that covers all the concepts and their differences as defined in equations (1) to (5), (14) and (17). Observing that (14) is the general case for n surveys of the particular case (7a) where $n = 2$, it becomes evident, that on taking differentials, squaring, summing, and dividing the linear sum of the n terms of (14) there results n^2 terms of which there are n that are variances (times constants) of the sort $\frac{k^2\sigma^2}{P}$ and $\frac{n^2 - n}{2}$ are different terms each occurring twice that are covariances (times constants) of the sort $\frac{kkQ\sigma\sigma r}{PP}$. From these

rough considerations as well as from rigorous derivation, the generalized standard error of (14) is found to be:

$$\sigma_f^2 = \sum_1^{n^2} \frac{k_{fi}\,\sigma_i\,k_{fj}\,\sigma_j\,Q_{ij}\,r_{ij}}{P_i\,P_j}, \tag{18}$$

The generalized standard error.

Where i and j denote each of the n surveys in turn. There will thus be n^2 terms to be summed—the number of combinations of i with j including the cases where $i = j$.

The derivation of (18) as well as its computation from data and its interpretation in special cases can all be made clearer by arranging the terms in a square array as follows:

	$i \rightarrow$	1	2	.....	n
j ↓	Coefficients ↓ →	$\frac{k_{f1}\sigma_1}{P_1}$	$\frac{k_{f2}\sigma_2}{P_2}$	.....	$\frac{k_{fn}\sigma_n}{P_n}$
1	$\frac{k_{f1}\sigma_1}{P_1}$	P_1 ( )	$Q_{12}r_{12}$ ( )	.....	$Q_{1n}r_{1n}$ ( )
2	$\frac{k_{f2}\sigma_2}{P_2}$	$Q_{12}r_{12}$ ( )	P_2 ( )	.....	$Q_{2n}r_{2n}$ ( )
⋮	⋮	⋮	⋮	.....	⋮
n	$\frac{k_{fn}\sigma_n}{P_n}$	$Q_{1n}r_{1n}$ ( )	$Q_{2n}r_{2n}$ ( )	.....	P_n ( )

To get σ_f write the computed values of the coefficients $\frac{k\sigma}{P}$ as captions of rows and of columns and write each computed Qr value in its appropriate cell, noting that in the main diagonal cells the self-correlations are unities and the population common to both column and row surveys, Q_{ij} is the entire population of that survey as $Q_{ij} = P_i$ when $i = j$. Thus $Q_{11} = P_1$. Next in each cell's parenthesis enter the product of three factors, namely: a) the cell Qr term, b) the column coefficient, and c) the row coefficient. The sum of these products in the parentheses, n^2 in number, is σ_f^2 of (18).

From the above square array it becomes clear that whenever in (17) the difference of two observed forces, or other functions, is derived from *different* populations the Q between these populations is zero so that the entire product terms in those cells vanish. Thus in the very simplest and familiar case of

comparing two means from different populations, $n = 2$, $Q_{12} = 0$, $k = 1$, and (18) reduces to the usual sum of the two variances of the two means

$$\sigma^2 \text{ difference in means} = \frac{\sigma_1^2}{P_1} + \frac{\sigma_2^2}{P_2} \tag{19}$$

IV. Some special cases

It should be observed that the above formulae for the standard errors when P shifts all become identical with the simpler formulae previously derived for the case of a constant P. In this case, every $Q_{pq} = P_p = P_q$ and in the square array (in addition to k's which no longer involve an average P), the Q or P of the cells and the P's in the row coefficients, may be omitted as they cancel each other out.

Another special but very frequent case is where the social change is not given in terms of a difference in means, S_1 and S_2, but in terms of a difference in percentages, as when a literacy rate rises from 30% to 40%. A percentage can be viewed as a mean of a two-category, all-or-none, present-or-absent variable such as: A, non-A (foreign or native born, literate or illiterate, etc), where A is assigned a value of 1 and non-A a value of 0. Then the sum of the values of A, each times its frequency, divided by the population is both a proportion and a mean. Its standard error in the percentage, p, form of expression is then equal to it in the mean form:

$$\sigma_p = \frac{p\sqrt{1.00 - p}}{\sqrt{P}} = \sigma_s = \frac{\sigma_s}{\sqrt{P}}$$
$$\left(\text{where } s = 1 \text{ or } 0 \text{ and } p = \frac{\Sigma s}{P} = S\right) \tag{20}$$

so that where S_i in (14) is a percent $p(1.00 - p)$ should be substituted for σ_i (and σ_j) in (18). In this case the appropriate formula to use for getting r_{ij} in (18) depends on the nature of the distribution of the variable that is expressed in percentage form. If the distribution is normal, tetrachoric r may be appropriate, while if the S in percentage form is from a two point distribution, r from a four fold point surface may be appropriate.

In all the above cases the usual interpretation of the significance of f in respect to sampling errors may be used in entering a normal probability table with a given σ_f from (18) and reading the probability of such a f occurring by chance.[4]

For a numerical illustration of this formula (18), consider the case of two villages, the statistical significance of whose momentums of a social change are to be determined. The data are from a study[1] of Syrian villages where an

[4] Mr. Wilks comments here that, "there is a more exact and rigorous test for comparing the two sets of S's which enter into a pair of M's or F's which involves some recent statistical theory but it is doubtful if the extra refinement is worth while at this stage of sociometric development."

itinerant Health Clinic in two years changed the average hygienic status of the families in each village by amounts of score (on a scale of 1 to 1000 points, devised for this study) as indicated in the table below.

	Village A	Village B
Mean score in 1931 = S_1 =	253	321
" " " 1933 = S_2 =	304	528
Population (families) in 1931 = P_1 =	46	46
" " " 1933 = P_2 =	40	32
Standard deviation of scores in 1931 = σ_1 =	54	39
" " " " " 1933 = σ_2 =	58	70
Families common to both censuses = Q_{12} =	40	32
Correlation of scores from the 2 dates = r_{12} =	.00	.19
$k_{M1} = -(P_1 + P_2)/2Y_{(2-1)}$ =	−21.5	−19.5
$k_{M2} = -k_{M1}$ =	21.5	19.5
$k_{M1}\sigma_1/P_1$ =	−25.24	−16.53
$k_{M2}\sigma_2/P_2$ =	31.17	42.65
$Q_{12}r_{12}$ =	0	6.08
$\sigma_{M'_{2-1}}$ =	261	249*
Momentum = M'_{2-1}	1,097	4,037
Significance ratio $M'_{2-1}/\sigma_{M'_{2-1}}$	4.2	16.2

* The calculation of this σ by (18) may be illustrated in detail:

Village B

Coefficients, $\frac{k\sigma}{P} \rightarrow$ / $\downarrow$		1	2	
		−16.53	42.65	$\Sigma(\ ) = 62{,}207$
1	−16.53	46 (= P_1) (12,571)	6.08 (= Qr) (−4,286)	$= \sigma^2_{M'_{(2-1)}}$
2	42.65	6.08 (= Qr) (−4,286)	32 (= P_2) (58,208)	$\sigma_{M'_{(2-1)}} = 249$

The momentum of the movement towards improved hygiene achieved in village A is 4.2 times its standard error, while that of village B is 16.2 times its standard error. The excess momentum of village A over village B is 8.1 $\left(= \frac{2940}{361}\right)$ times the standard error of their difference in momenta. Since all three of these significance ratios are well over 3 the conclusion is that the observed momenta and difference of momenta are statistically significant and cannot reasonably be due to sampling fluctuations. It may be noted that the significance ratios for the amounts of this social *change*, the difference in mean scores, are in close agreement with the above figures, being 4.1 and 15.9 for

villages A and B respectively, instead of 4.2 and 16.2 as above. These discrepancies of a .1 and .3 in the statistical significance of these social changes compared with the corresponding social momenta are accounted for by the fact that the shift in the size of the population is allowed for in our formula for the case of momenta and is not considered in the usual formula for the case of social change.

A minimum of three measurements of one population is necessary to determine a social force. To determine its standard error all the correlations must be secured between every pair of measurements, each correlation derived from the part of the total population that is common to that pair of measurements. Obviously the data as currently reported from surveys and censuses and statistical bureaus do not meet these specifications. More rigorous analysis of social data and reporting of correlations in it is a prerequisite to the measurement of social forces and their significance.

AMERICAN UNIVERSITY OF BEIRUT, SYRIA.

AN APPROXIMATION TO "STUDENT'S" DISTRIBUTION*

By Walter A. Hendricks

I. Introduction

The function commonly known as "Student's" distribution occupies a prominent position among the classic contributions to the field of statistics, not only for its intrinsic value but also for the stimulus which it gave to statistical research at the time of its discovery.

The function, which may be written in the form,

$$dF_z = \frac{1}{B[\frac{1}{2}(n-1), \frac{1}{2}]} (1 + z^2)^{-\frac{1}{2}n} dz, \tag{1}$$

gives the distribution of the ratio, z, of the estimated arithmetic mean, $\bar{x}$, to the estimated standard deviation, s, for samples of n observations drawn from the normal universe specified by the arithmetic mean, zero, and the standard deviation, σ. This function, together with a table of values of its integral was given by "Student."[9, 10]

In view of the fact that similar distributions were subsequently found by Fisher[2] to arise in a larger variety of practical problems than was originally supposed, a table of values of a new integral was later given by "Student"[11] in which the distribution of a variable, t, defined by the relation,

$$t = (n-1)^{\frac{1}{2}} z, \tag{2}$$

rather than the distribution of z itself, was considered. Another table giving the distribution of t, in a form intended to be more convenient for use by research workers wishing to apply statistical methods to experimental data, was later given by Fisher.[3]

The integration of functions of the type defined by equation (1) involves considerable labor, a fact which has been somewhat embarrassing to practical statisticians interested in the distributions of z and t for values of n larger than those included in the above-mentioned tables. The recent appearance of Tables of the Incomplete Beta-Function, prepared under the direction of Pearson,[7] has considerably alleviated the difficulty, but the requirements of certain practical problems are not easily satisfied even with the aid of these tables. Consequently, simple approximations to the distributions of z and t,

*A thesis submitted to the Faculty of the Columbian College of The George Washington University in part satisfaction of the requirements for the degree of Master of Arts.

which will be sufficiently accurate for most practical purposes, should be of some interest.

According to "Student,"[9] the distribution of z tends to approach a normal curve with a standard deviation of $(n-3)^{-\frac{1}{2}}$ for values of n greater than 10. However, Deming and Birge[1] have recently suggested that the distribution tends to approach a normal curve with a standard deviation of $(n-1\frac{1}{2})^{-\frac{1}{2}}$.

This thesis presents a simple approximation to the distribution of z, which can be readily extended to the distribution of t and which will give more accurate results than either of the above approximations.

II. Approximation to the Distribution of Z

The approximation presented here is based upon the assumption that, for large values of n, the distribution of s tends to approach a normal curve with the arithmetic mean, $\bar{s}$, and the standard deviation, $\frac{\sigma}{2^{\frac{1}{2}}n^{\frac{1}{2}}}$, that is,

$$dF_s = \frac{n^{\frac{1}{2}}}{\pi^{\frac{1}{2}}\sigma} e^{-\frac{n}{\sigma^2}(s-\bar{s})^2} ds. \tag{3}$$

Since the distribution of the estimated arithmetic mean, $\bar{x}$, is known to be normal, with the standard deviation, $\frac{\sigma}{n^{\frac{1}{2}}}$, we have for the joint distribution of s and $\bar{x}$:

$$dF_{s,\bar{x}} = \frac{n}{2^{\frac{1}{2}}\pi\sigma^2} e^{-\frac{n}{\sigma^2}[\frac{1}{2}\bar{x}^2+(s-\bar{s})^2]} ds\, d\bar{x}. \tag{4}$$

$\bar{s}$ may be expressed in terms of n and σ by the well-known relation,

$$\bar{s} = c_n\sigma, \tag{5}$$

in which the factor, c_n, is defined by the formula,

$$c_n = \frac{2^{\frac{1}{2}}}{n^{\frac{1}{2}}} \frac{\Gamma(\frac{1}{2}n)}{\Gamma[\frac{1}{2}(n-1)]}. \tag{6}$$

If we write, $c_n\sigma$, in place of $\bar{s}$, in equation (4) and make the transformation,

$$\bar{x} = sz, \tag{7}$$

we have for the joint distribution of s and z:

$$dF_{s,z} = \frac{n}{2^{\frac{1}{2}}\pi\sigma^2} e^{-\frac{n}{\sigma^2}[\frac{1}{2}s^2z^2+(s-c_n\sigma)^2]} s\, ds\, dz. \tag{8}$$

To find the distribution of z, all that is necessary is to write:

$$dF_z = k\left[\int_{-\infty}^{+\infty} e^{-(as-b)^2} s\, ds\right] dz, \tag{9}$$

in which:

$$k = \frac{n}{2^{\frac{1}{2}}\pi\sigma^2} e^{-nc_n^2 \frac{z^2}{z^2+2}},$$

$$a = \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\sigma} (z^2 + 2)^{\frac{1}{2}}, \tag{10}$$

$$b = 2^{\frac{1}{2}} n^{\frac{1}{2}} c_n (z^2 + 2)^{-\frac{1}{2}}.$$

The integral in brackets in equation (9) can be evaluated without any difficulty. We have:

$$\int_{-\infty}^{+\infty} e^{-(as-b)^2} s\, ds = \frac{b\pi^{\frac{1}{2}}}{a^2}. \tag{11}$$

Substituting this value in equation (9) and replacing k, a, and b by the quantities which they represent, we obtain the following expression for the distribution of z:

$$dF_z = \frac{2n^{\frac{1}{2}}c_n}{\pi^{\frac{1}{2}}} e^{-nc_n^2 \frac{z^2}{z^2+2}} (z^2 + 2)^{-\frac{1}{2}}\, dz. \tag{12}$$

If we now define a new variable, u, by the relation,

$$u^2 = 2nc_n^2 \frac{z^2}{z^2 + 2}, \tag{13}$$

and make the appropriate substitutions in equation (12), we have, for the distribution function of u:

$$dF_u = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} e^{-\frac{1}{2}u^2}\, du. \tag{14}$$

Equation (14) is obviously a normal curve with unit standard deviation. We have thus deduced the interesting fact that, for values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantity, $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2 + 2)^{\frac{1}{2}}}$, is distributed as a normal deviate with unit standard deviation.

The accuracy of this approximation as compared with that of the approximation suggested by "Student"[9] and that of the more recent approximation suggested by Deming and Birge[1] may now be considered. As previously stated, the "Student" approximation is based on the assumption that the quantity, $(n - 3)^{\frac{1}{2}}z$, is distributed as a normal deviate with unit standard deviation for values of n greater than 10, while that suggested by Deming and Birge is based on the assumption that the quantity, $(n - 1\frac{1}{2})^{\frac{1}{2}}z$, is so distributed.

Table 1* gives values of the integral, I_z, defined by:

$$I_z = \frac{1}{B[\frac{1}{2}(n-1), \frac{1}{2}]} \int_{-\infty}^{z} (1 + z^2)^{-\frac{1}{2}n}\, dz, \tag{15}$$

* All tables and charts to which reference is made are to be found in the Appendix.

for the case, $n = 10$, together with the corresponding approximate values obtained by making use of the three approximations suggested by "Student," Deming and Birge, and the present author, respectively. The exact values and those obtained by the "Student" approximation were derived from values calculated by "Student"[9] and given by Pearson.[5] All other data in the table were calculated by the present author.

An inspection of Table 1 shows that the values of I_z based on the approximation presented in this thesis agree very well with the corresponding exact values. The agreement is better than that found in the case of either of the other two approximations. The Deming and Birge approximation gives better results than the "Student" approximation for values of z in the neighborhood of zero, but for other values of z the opposite is true.

III. Approximation to the Distribution of t

Since tables giving the distribution of the variable, t, have largely superseded those giving the distribution of z in practical statistical work, the feasibility of applying the above three approximations to the distribution of t is worthy of consideration.

The variable, t, has already been defined in terms of n and z by equation (2). If, in equation (12), we make the transformation,

$$z = (n-1)^{\frac{1}{2}}t, \tag{16}$$

we have, for the distribution function of t:

$$dF_t = \frac{2n^{\frac{1}{2}}(n-1)c_n}{\pi^{\frac{1}{2}}} e^{-nc_n^2 \frac{t^2}{t^2+2(n-1)}} [t^2 + 2(n-1)]^{-\frac{1}{2}}\, dt. \tag{17}$$

If we now define a variable, v, by the relation,

$$v^2 = 2nc_n^2 \frac{t^2}{t^2 + 2(n-1)}, \tag{18}$$

we have, for the distribution function of v:

$$dF_v = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} e^{-\frac{1}{2}v^2}\, dv. \tag{19}$$

Equation (19) shows that, for values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantity,

$$2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{t}{[t^2 + 2(n-1)]^{\frac{1}{2}}},$$

is distributed as a normal deviate with unit standard deviation. On the other hand, if we assume with "Student" that, for large values of n, the quantity, $(n-3)^{\frac{1}{2}}z$, is normally distributed about zero with unit standard deviation, we should expect to find that the quantity, $\frac{(n-3)^{\frac{1}{2}}}{(n-1)^{\frac{1}{2}}}t$, is also distributed as a normal

deviate with unit standard deviation. If the Deming and Birge approximation to the distribution of z is assumed to be valid, we should expect to find that the quantity, $\frac{(n-1\frac{1}{2})^{\frac{1}{2}}}{(n-1)^{\frac{1}{2}}}t$, is distributed as a normal deviate with unit standard deviation.

To test the accuracy of each of these three approximations to the distribution of t, we may make use of the well-known table of values of t given by Fisher.[3] This table is so constructed that a value of t corresponding to a given number of "degrees of freedom" and a given value of "P" may be read from the table, where P is defined by the relation,

$$P = 1 - \frac{2}{(n-1)^{\frac{1}{2}}B[\frac{1}{2}(n-1), \frac{1}{2}]}\int_0^t \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n} dt. \tag{20}$$

The entries in the last line of the table, corresponding to an infinite number of "degrees of freedom," are the deviates of a normal curve with unit standard deviation.

To test the accuracy of the "Student" approximation, we may calculate the entries for a line of this table, corresponding to $n - 1$ "degrees of freedom," by multiplying the entries in the last line of the table by $\frac{(n-1)^{\frac{1}{2}}}{(n-3)^{\frac{1}{2}}}$. These approximate values of t may then be compared with the exact values given in the table. The accuracy of the Deming and Birge approximation may be tested in the same manner, except that in this case the entries in the last line of the table should be multiplied by $\frac{(n-1)^{\frac{1}{2}}}{(n-1\frac{1}{2})^{\frac{1}{2}}}$. To test the accuracy of the approximation given by equation (19), we may calculate the values of t corresponding to $n - 1$ "degrees of freedom" by means of the relation,

$$t^2 = \frac{2(n'-1)v^2}{2nc_n^2 - v^2}, \tag{21}$$

in which the entries in the last line of the table are to be taken as the values of v.

Table 2 gives the exact values of t corresponding to the values of P given in Fisher's table for $n = 10$, together with the approximate values calculated by means of each of the above three approximations. This comparison of the accuracies of the three approximations is equivalent to the comparisons presented in Table 1. The conclusions which may be drawn are in agreement with those which have already been drawn from that table.

In order to test the behavior of each of the approximations for a larger value of n, values of t corresponding to the different values of P were calculated for $n = 30$. The results are presented in Table 3. The rank of each of the three approximations, with regard to accuracy, for $n = 30$ is the same as for $n = 10$. Although all three give more accurate results for the larger value of n, the superiority of the approximation presented in this thesis is quite apparent.

For extremely large values of n, all three approximations evidently tend to become one-hundred percent accurate, for the distribution of t tends to become normal as n is increased indefinitely. In the case of the "Student" and Deming and Birge approximations, the ratios, $\frac{(n-1)^{\frac{1}{2}}}{(n-3)^{\frac{1}{2}}}$ and $\frac{(n-1)^{\frac{1}{2}}}{(n-1\frac{1}{2})^{\frac{1}{2}}}$, obviously approach unity, respectively, as n becomes very large. The approximate value of t given by equation (21) also tends to approach the normal deviate, v, as n is increased for we have:

$$\lim_{n\to\infty} t^2 = \lim_{n\to\infty} \frac{2(n-1)v^2}{2nc_n^2 - v^2} = \lim_{n\to\infty} \left[\frac{2nv^2}{2nc_n^2 - v^2} - \frac{2v^2}{2nc_n^2 - v^2}\right]$$

(22)

$$= \lim_{n\to\infty} \left[\frac{v^2}{c_n^2 - \frac{v^2}{2n}} - \frac{2v^2}{2nc_n^2 - v^2}\right] = v^2.$$

IV. Discussion

The greater accuracy of the approximation to the distribution of z presented in this thesis apparently can not be explained by the hypothesis that the distribution of s becomes normal more rapidly than the distribution of z as n is increased. Table 4 presents values of the ordinates of the normal curve with unit standard deviation, together with the corresponding ordinates of the exact distributions of the quantities, $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$, $(n-3)^{\frac{1}{2}}z$, $(n-1\frac{1}{2})^{\frac{1}{2}}z$, and $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2+2)^{\frac{1}{2}}}$, for $n = 10$. Although the distribution of $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$ seems to follow the normal curve more closely than does the distribution of $(n-3)^{\frac{1}{2}}z$, the opposite seems to be true in the case of the distribution of $(n-1\frac{1}{2})^{\frac{1}{2}}z$. The distribution of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2+2)^{\frac{1}{2}}}$, however, follows the normal curve quite closely.

The behavior of these distributions for $n = 10$ can be observed more easily in Figures 1, 2, and 3 in which the frequency curves of $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$, $(n-3)^{\frac{1}{2}}z$, and $(n-1\frac{1}{2})^{\frac{1}{2}}z$ are respectively plotted together with the normal curve with unit standard deviation. The frequency curve of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2+2)^{\frac{1}{2}}}$ was not plotted because of the fact that this curve follows the normal curve so closely that the two curves could not be distinguished when plotted on the scale used in the other three charts.

The most reasonable conclusion which can be drawn from Table 4 and Figures 1, 2, and 3 is that the departure of the exact distribution of s from the normal curve has very little effect in destroying the normality of the distribution of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2+2)^{\frac{1}{2}}}$.

V. Values of the Factor, c_n

For the practical application of the approximations to the distributions of z and t presented in this thesis, a table of values of the factor, c_n, is required. Values of this factor, for values of n as high as 100, have been tabulated by Pearson[4, 6] and by Shewhart.[8] For values of n greater than 100, c_n may be calculated accurately to at least five significant figures by the following relation, given by Pearson[4] and by Deming and Birge[1]:

$$c_n = 1 - \frac{3}{4n} - \frac{7}{32n^2}. \tag{23}$$

Table 5 presents values of c_n for some large values of n, calculated by the present author. For values of n not included in this table, c_n may be calculated by means of equation (23) just as rapidly as by interpolation in the table.

VI. Summary and Conclusions

For values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantities,

$$2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2+2)^{\frac{1}{2}}} \text{ and } 2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{t}{[t^2+2(n-1)]^{\frac{1}{2}}},$$

are distributed as normal deviates with unit standard deviation. The results obtained by assuming a normal distribution of s are more accurate than those obtained by assuming that either $(n-3)^{\frac{1}{2}}z$ or $(n-1\frac{1}{2})^{\frac{1}{2}}z$ is distributed as a normal deviate with unit standard deviation. For extremely large values of n, the distribution of each of the above quantities tends to approach a normal curve with a mean of zero and unit standard deviation.

VII. References

(1) Deming, W. Edwards, and R. T. Birge, 1934. On the statistical theory of errors. Reviews of Modern Physics, 6: 119–161.
(2) Fisher, R. A., 1925. Applications of "Student's" distribution. Metron, 5: 90–104.
(3) Fisher, R. A., 1934. Statistical Methods for Research Workers, 5th ed. Oliver and Boyd, Edinburgh and London.
(4) Pearson, Karl, 1915. On the distribution of the standard deviations of small samples. Biometrika, 10: 522–529.
(5) Pearson, Karl, 1924. Tables For Statisticians And Biometricians, Part I, 2nd ed. Cambridge University Press, Cambridge.
(6) Pearson, Karl, 1931. Tables For Statisticians And Biometricians, Part II, 2nd ed. Cambridge University Press, Cambridge.
(7) Pearson, Karl, 1934. Tables Of The Incomplete Beta-Function. Cambridge University Press, Cambridge.
(8) Shewhart, W. A., 1931. Economic Control Of Quality Of Manufactured Product. D. Van Nostrand Co., New York.
(9) "Student," 1908. The probable error of a mean. Biometrika, 6: 1–25.
(10) "Student," 1917. Tables for estimating the probability that the mean of a unique sample of observations lies between $-\infty$ and any given distance of the mean of the population from which the sample is drawn. Biometrika, 11: 414–417.

(11) "Student," 1925. New tables for testing the significance of observations. Metron, 5:105-120.

THE GEORGE WASHINGTON UNIVERSITY.

VIII. Appendix

TABLE 1

Exact values of I_z and approximate values, derived from tables of the normal probability integral, for $n = 10$

z	I_z			
	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
−2.0	.0001	.0000	.0000	.0004
−1.8	.0002	.0000	.0000	.0006
−1.6	.0005	.0000	.0000	.0010
−1.4	.0011	.0001	.0000	.0018
−1.2	.0029	.0007	.0002	.0038
−1.0	0075	.0041	.0018	.0086
− .8	.0199	.0171	.0098	.0211
− .6	.0527	.0562	.0401	.0535
− .4	.1304	.1448	.1218	.1307
− .2	.2816	.2984	.2799	.2817
.0	.5000	.5000	.5000	.5000
+ .2	.7184	.7016	.7201	.7183
+ .4	.8696	.8552	.8782	.8693
+ .6	.9473	.9438	.9599	.9465
+ .8	.9801	.9829	.9902	.9789
+1.0	.9925	.9959	.9982	.9914
+1.2	.9971	.9993	.9998	.9962
+1.4	.9989	.9999	1.0000	.9982
+1.6	.9995	1.0000	1.0000	.9990
+1.8	.9998	1.0000	1.0000	.9994
+2.0	.9999	1.0000	1.0000	.9996

TABLE 2

Exact values of t corresponding to different values of P and approximate values, derived from normal deviates, for n = 10

P	t			
	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
.90	.129	.142	.129	.129
.80	.261	.287	.261	.261
.70	.398	.437	.396	.398
.60	.543	.595	.540	.544
.50	.703	.765	.694	.703
.40	.883	.954	.866	.884
.30	1.100	1.175	1.066	1.104
.20	1.383	1.453	1.319	1.386
.10	1.833	1.865	1.693	1.844
.05	2.262	2.222	2.017	2.290
.02	2.821	2.638	2.394	2.896
.01	3.250	2.921	2.650	3.389

TABLE 3

Exact values of t corresponding to different values of P and approximate values, derived from normal deviates, for n = 30

P	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
.90	.127	.130	.127	.127
.80	.256	.263	.256	.256
.70	.389	.399	.389	.389
.60	.530	.543	.529	.530
.50	.683	.699	.680	.683
.40	.854	.872	.849	.854
.30	1.055	1.074	1.045	1.055
.20	1.311	1.328	1.293	1.312
.10	1.699	1.705	1.659	1.700
.05	2.045	2.031	1.977	2.047
.02	2.462	2.411	2.347	2.466
.01	2.756	2.670	2.598	2.764

TABLE 4

Ordinates of the normal curve with unit standard deviation and ordinates of the exact distribution functions of $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s-\bar{s})$, $(n-3)^{\frac{1}{2}}z$, $(n-1\frac{1}{2})^{\frac{1}{2}}z$, *and* $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\frac{z}{(z^2+2)^{\frac{1}{2}}}$ *for* $n = 10$

Deviation from mean	Ordinates of distribution function				
	Normal deviate	$\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s-\bar{s})$	$(n-3)^{\frac{1}{2}}z$	$(n-1\frac{1}{2})^{\frac{1}{2}}z$	$2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\frac{z}{(z^2+2)^{\frac{1}{2}}}$
−3.0	.0044	.0006	.0071	.0108	.0034
−2.5	.0175	.0085	.0181	.0254	.0156
−2.0	.0540	.0454	.0459	.0581	.0544
−1.5	.1295	.1356	.1092	.1234	.1306
−1.0	.2420	.2663	.2256	.2290	.2426
− .5	.3521	.3751	.3692	.3454	.3522
.0	.3989	.3999	.4400	.3991	.3990
+ .5	.3521	.3343	.3692	.3454	.3522
+1.0	.2420	.2245	.2256	.2290	.2426
+1.5	.1295	.1233	.1092	.1234	.1306
+2.0	.0540	.0560	.0459	.0581	.0544
+2.5	.0175	.0213	.0181	.0254	.0156
+3.0	.0044	.0068	.0071	.0108	.0034

TABLE 5

Values of c_n *for large values of* n

n	c_n	n	c_n
100	.99248	900	.99917
150	.99499	1000	.99925
200	.99624	2000	.99962
250	.99700	3000	.99975
300	.99750	4000	.99981
350	.99786	5000	.99985
400	.99812	10000	.99992
450	.99833	20000	.99996
500	.99850	30000	.99997
600	.99875	40000	.99998
700	.99893	50000	.99998
800	.99906	100000	.99999

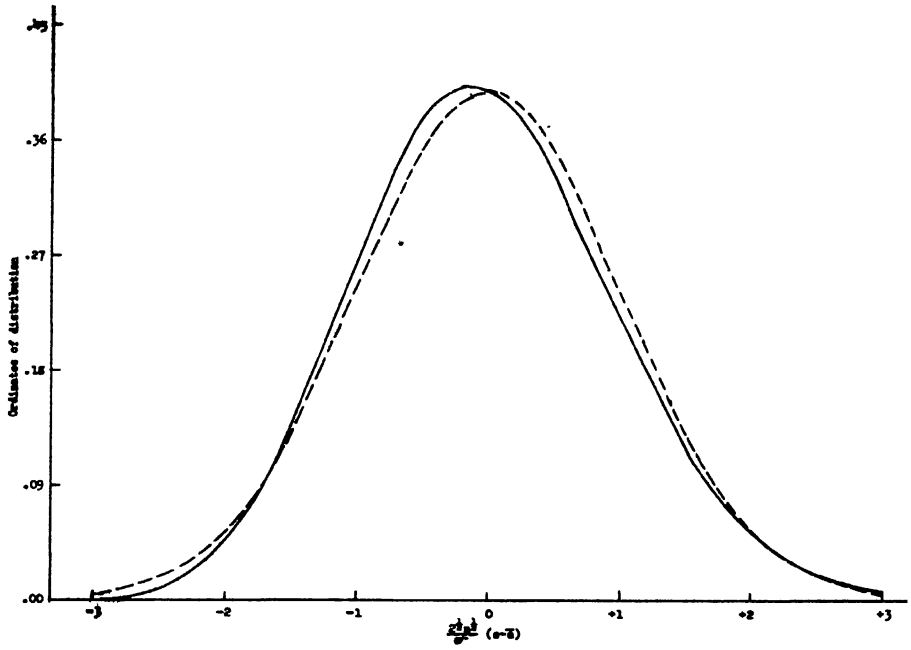


FIG. 1. EXACT DISTRIBUTION OF $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s-\bar{s})$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

————, exact distribution; ----------, normal curve

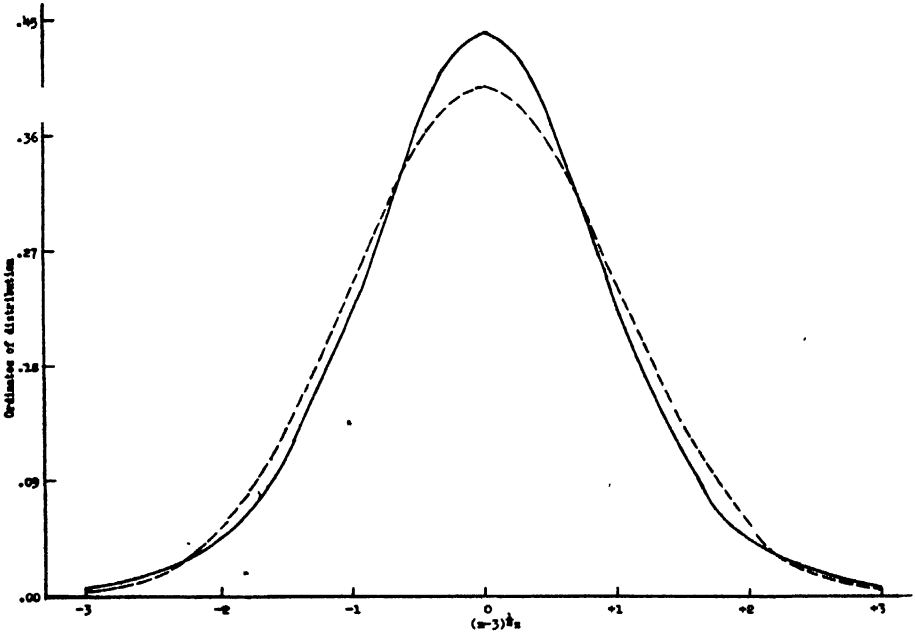


FIG. 2. EXACT DISTRIBUTION OF $(n-3)^{\frac{1}{2}}z$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

————, exact distribution; ----------, normal curve

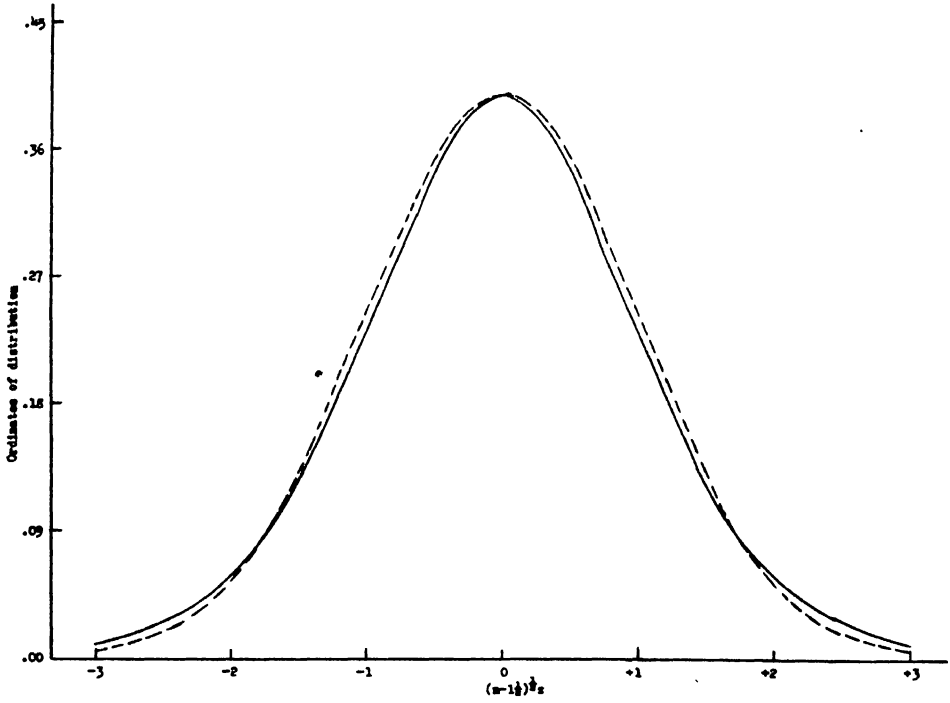


FIG. 3. EXACT DISTRIBUTION OF $(n - 1\frac{1}{2})^{\frac{1}{2}}z$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

————, exact distribution; ----------, normal curve